This is to certify that the
dissertation entitled

AN INVESTIGATION OF USING COLLATERAL INFORMATION TO REDUCE EQUATING BIASES OF THE POST-STRATIFICATION EQUATING METHOD

presented by

Sungworn Ngudgratoke

has been accepted towards fulfillment
of the requirements for the

Ph.D. degree in Measurement and Quantitative Methods

Mark D. Reckase
Major Professor's Signature

July 28, 2009
Date

AN INVESTIGATION OF USING COLLATERAL INFORMATION TO REDUCE EQUATING BIASES OF THE POST-STRATIFICATION EQUATING METHOD

By

Sungworn Ngudgratoke

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Measurement and Quantitative Methods

2009

ABSTRACT

AN INVESTIGATION OF USING COLLATERAL INFORMATION TO REDUCE EQUATING BIASES OF THE POST-STRATIFICATION EQUATING METHOD

By

Sungworn Ngudgratoke

In many educational assessment programs, the use of multiple test forms developed from the same test specification is very common because requiring different examinees to take different test forms of the same test makes it possible to maintain the security of the test. When multiple test forms are used, it is necessary to make the assessment fair to all examinees by using a statistical procedure called "equating" to adjust for differences in the test forms. If equating is successfully carried out, equated scores are comparable as if they were from the same test form.

Two commonly used observed score equating methods that use the Non-Equivalent groups with Anchor Test (NEAT) design to collect equating data include the chain equating (CE) method and the post-stratification equating (PSE) method. It has been documented that the CE method produced smaller equating biases than the PSE method, when two groups of examinees differ greatly in abilities. Therefore, the CE method has been used more widely in practice, even though the PSE method is more theoretically sound than the CE method. Larger equating biases are due to the fact that the anchor test score fails to remove unintended differences between groups of examinees.

Aiming to reduce equating biases of the PSE method, this study used collateral information about examinees as a new way to construct synthetic population functions, rather than a single variable such as the anchor test score or the anchor test true score. Collateral information used in this study included the anchor test score, sub-scores, and examinees' demographic variables. This study investigated two different methods of using such collateral information about examinees to improve equating results of the PSE method. These two methods included the propensity score method (Rosenbaum & Rubin, 1983) and the multiple imputation method (Rubin, 1987). Both simulation data and empirical data were used to develop the equating function to explore if it was feasible to use collateral information to reduce equating biases under different conditions including test length, group differences, and missing data treatment.

The results from simulation data show that sub-scores or sub-scores combined with other collateral information in a form of propensity scores had a potential to reduce equating biases for long tests, when there were group differences in abilities. However, demographic variables had a potential to reduce equating biases for the multiple imputation method.

ACKNOWLEDGEMENTS

My deepest gratitude goes to Dr. Mark D. Reckase. His wisdom inspired me to come to Michigan State University to pursue my doctoral degree. Without his guidance, support, and insightful comments, this work would not have been possible. I would like to thank members of my dissertation committee: Dr. Richard T. Houang, Dr. Kimberly S. Maier, and Dr. Alexander Von Eye for their helpful suggestions on this study.

I also would like to thank the Royal Thai government for giving me financial support for my graduate study at MSU. I would like to thank Mike Sherry and Dipendra Subedi for their comments on early versions of my dissertation.

Finally, I wish to thank my parents, grandmother, brother and sister for their love, patience, and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

Equating is a statistical procedure that is used to adjust scores on test forms (e. g., X and Y) so that scores on the forms can be used interchangeably (Kolen & Brennan, 2004) as if scores are from the same test forms. Without equating, examinees taking the easier test form will have an unfair advantage. When test score equating is performed, standard errors of equating should be estimated to quantify equating errors (AERA, APA, NCME, 1999). Accurate equating results not only facilitate test score interpretations but also enhance fair comparisons across individuals, states, and countries.

The Non-Equivalent Groups with Anchor Test (NEAT) design is commonly used in equating practice because the anchor test score (A) can adjust for preexisting differences between examinees. In this design, two operational tests to be equated, X and Y, are given to two samples of examinees from potentially different test populations (referred to as P and Q). In addition, an anchor test, A, is given to both samples from P and Q. In observed score equating, when equating data are collected through the NEAT design, test score equating can be performed using a variety of observed score equating methods such as the Tucker method, the Levine observed score method, the Levine true score method, the post-stratification method, the Braun and Holland linear method, and the chain equipercentile method. Among these methods, the post-stratification equating (PSE) method (von Davier, Holland & Thayer, 2004) which is also called the "frequency estimation method" (Kolen & Brennan, 2004) and the chain equipercentile (CE) method are two important methods commonly used in practice (Holland, Sinharay, von Davier, &

Han, 2007). The role of an anchor test (A) in the PSE method is not only to remove differences between P and Q but also to estimate score frequencies of the designed missing data in the NEAT design so that synthetic population functions required to derive comparable scores using the equipercentile equating function can be constructed (Braun & Holland, 1982).

The PSE method is based on a strong theoretical foundation that centers on the generalization of equating function linking X-scores to Y-scores (Harris & Kolen, 1990; Kolen, 1992), making it more appealing than the CE method. More specifically, the equating function is computed for a single population. When P and Q differ greatly in abilities, we do not know for what population the equating function is computed. The PSE method defines the target population in the form of synthetic population functions (T) which are mixtures of both P and Q. In contrast, the CE method is considered to be less sound in its theoretical foundation because it does not define any synthetic population functions.

Even though the PSE is more theoretically sound than the CE method, it produces unfavorable equating functions and researchers prefer the CE method to the PSE method. Braun and Holland (1982) noted that the PSE and CE methods give different results. The CE is more preferable than the PSE method since it produces smaller equating bias. Holland, Sinharay, von Davier, and Han, (2007); and Wang, Lee, Brennan, and Kolen (2008) found that when groups differ in abilities, the PSE method produces larger equating bias but less standard errors of equating than does the CE method. This might be because the anchor test fails to remove the bias to which the nonequivalence of P and Q

can lead. The bias due to preexisting differences between groups that cannot be removed by the anchor test precludes valid interpretations and fair uses of test scores.

To reduce equating bias when P and Q differ greatly in abilities, the propensity score (Rosenbaum & Rubin, 1983) can be a desirable method to augment the PSE method (Livingston, Wright, & Dorans, 1993). In the equating context, the propensity score is the estimated conditional probability that a subject will be assigned to a particular test form, given a vector of observed covariates (e.g., demographic variables and anchor test scores). Covariates used to estimate examinees' propensity scores are called "collateral information", which is available information about examinees in addition to their item responses (Mislevy, Kathleen, & Sheehan, 1989). Any examinees with equal propensity scores are homogeneous in terms of covariates. Propensity scores computed from both demographic variables and anchor test scores may be intuitively advantaged because score frequencies of missing data may be better estimated and thus synthetic population functions might be precisely estimated, resulting in the more accurate equating function. Because more covariates have a potential to handle missing data in the NEAT design, less equating bias is expected. Therefore, the propensity score method may be another equating method alternative to the method that uses a special anchor test (Sinharay & Holland, 2007) and the method that uses the anchor test true score (Wang & Brennan, 2009). The Sinharay and Holland's method uses an anchor test composed of a large number of medium difficulty items, and it is appropriate for equating that uses an external anchor test only because the special anchor test construction may not meet the test specification well (Sinharay & Holland, 2007).

However, it is found that when groups differ greatly, using a few demographic variables in combination with anchor test scores does not improve equating accuracies (Paek, Liu, & Oh, 2008). Therefore, it is necessary to find more collateral information about examinees that is available from the test to adjust for group differences. The variable that is promising in this regard is the subscore which is a score on the subsection of the test. For example, a test measuring mathematics proficiency may contain subsections such as algebra, functions, geometry, and number and operation, and a subscore is the score assigned to a subsection of the test. Subscore reporting usually provides more detailed diagnostic information about examinees' performance that may be useful, for example, in making individual instruction placement and remediation decisions (Tate, 2004) and in formatively supporting teaching and learning (Dibello & Stout, 2007). In equating, subscores are expected to give accurate equating functions since high correlations between operational test scores and scores on subtests make them feasible to compute missing data on the operational test through an existing missing data treatment method such as the multiple imputation method (e.g. Rubin, 1987; Schafer, 1997). Therefore the combination of anchor test scores, subscores, and demographic variables are worth investigating if it could improve the PSE equating method.

In this study, demographic variables, the anchor test scores, and subscores are called collateral information. To improve the accuracy of the PSE equating results, this study proposed using two different approaches of using this collateral information to equate test scores using the PSE method. These two approaches include the propensity scores (Rosenbaum & Rubin, 1983) and the multiple imputation method (Rubin, 1987; Schafer, 1997). These two methods were used to fill in missing data in the NEAT design.

For the propensity score method, demographic variables, subscores and the anchor test score were combined into examinees' propensity score with which the anchor test score will be replaced in the PSE method. For the multiple imputation method, demographic variables, subscores, and the anchor test scores were used as covariates to fill in missing data.

1.2 Research Questions

Using both real and simulated data, this study explored the feasibility of using combinations of demographic variables, anchor test scores, and subscores in two different ways to increase the precision of the PSE method. The main research question concerns the potential use of subscores combined with demographic variables and an anchor test score in improving the accuracy of the PSE equating results. In this study, the "traditional PSE method" refers to the PSE method that uses the anchor test score in the PSE equating process (von Davier, Holland, & Thayer, 2004). This method is the same as the frequency estimation method (Kolen & Brennan, 2004). The PSE method that replaces anchor test scores with the anchor test true score (Wang & Brennan, 2009) is called the "modified PSE method" in this study to contrast between the original PSE method and the modified PSE method, even though it is originally called the "modified frequency estimation method" by Wang and Brennan. The more specific research questions are as follows:

1. How accurate are predicted score frequencies of missing data when the proposed methods are used to compute missing data?
2. How comparable are predicted score frequencies of missing data produced by the proposed methods and those produced by the traditional and the modified PSE methods?

3. How do the proposed methods influence equating bias, and standard errors of equating?
4. How comparable are the proposed methods to the traditional and modified PSE methods in terms of equating bias, and standard errors of equating?

CHAPTER II

LITERATURE REVIEW

This study explored benefits of using collateral information as an alternative approach to construct synthetic population functions that are required for equating test scores using the post-stratification equating (PSE) method. The PSE method is an equating method that uses the non-equivalent groups with anchor test (NEAT) design. The PSE method and chain equipercentile equating (CE) method are two important equipercentile function-based equating methods that use the NEAT design. This study focuses on the PSE method only, because it is developed based on more sound theoretical foundation than the CE method. The improvement of the PSE is needed because it has been shown in equating literature that even though it is based on a strong theoretical foundation, it produced large equating biases than the CE method, especially when groups differ greatly in abilities. In this regard, this study used collateral information, as an alternative approach to improve equating results of the PSE method. This study explored the uses of collateral information the PSE method in two different ways, which will be explained in the next section.

This chapter reviews concepts and methodologies relating to equating and the development of the PSE method, and to the existing measurement and statistical developments that used to develop the approaches to enhancing PSE equating results in this study. Specifically, there are 12 related sections presented in this chapter which are outlined as follows:

Section 2.1 Collateral information

Section 2.2 General idea and the importance of test score equating.

Section 2.3 Equipercentile equating function

Section 2.4 Basic idea of the non-equivalent groups with anchor test design (NEAT).

Section 2.5 Importance of the synthetic population function

Section 2.6 Log-linear presmoothing technique and how it is important in test equating

Section 2.7 Post-stratification equating (PSE) method and synthetic population functions

Section 2.8 Method of sub-score estimation

Section 2.9 Propensity scores and the logistic regression approach to propensity score estimation.

Section 2.10 Multiple imputation method

Section 2.11 Statement of research problem

Section 2.12 Goal of this study and the evaluation indices

2.1 Collateral Information

In test score equating, only test scores are involved in the equating process. However, when groups of examinees differ greatly in terms of abilities, it is desire that collateral information about examinees be included in the equating process to reduce biases due to the group differences (Livingston, Dorans, Wright, 1990; Kolen, 1990). Collateral information is available information about examinees in addition to their item responses (Mislevy, Kathleen, & Sheehan, 1989). Familiar examples include demographic variables, and educational variables such as opportunity to learn variables

and grade received. Collateral information used in this study includes subscores, demographic variables, opportunity-to-learn variables, and the anchor test score.

2.2 Equating

The use of multiple test forms of the same test is a common practice by many large-scale assessment programs because of security issues. For example, administering different forms of the same test to different groups of test takers ensures that a large portion of items in the item bank will not be exposed to examinees. However, when multiple forms are used, it is possible that one test form could be harder than another. This existence of test forms with unequal difficulty raises a question regarding the fairness of testing which is a major concern for most testing programs. To make assessments fair to all examinees, it is therefore necessary to adjust for unintended difficulties that are left imbalanced across test forms by using a statistical adjustment called equating. Equating is a statistical method used to produce scores that can be used interchangeably (Kolen & Brennan, 2004) and one of its advantage is that when equating is successfully done, it produces comparable scores that are fair to test takers who take different test forms with unequal difficulty.

Equating and linking are two different terms that express how scores on different test forms are transformed to each other. Even though they have different meaning, both of these terms are best understood among other score transformation methods such as anchoring, calibration, statistical moderation, scaling, and prediction (Linn, 1996). While linking is a generic name for score transformation, equating is the most demanding type of linking (Linn, 1996). Transforming scores from one test form to scores on another test form cannot be called equating if it does not satisfy requirements of equating. That is,

tests to be equated should meet requirements as stated by Lord (1980) and by Doran and Holland (2000). Broadly speaking, tests being equated to each other should measure the same construct, have equal reliability, and also satisfy three requirements: symmetry, equity, and population invariance.

The Standard 4.11of *Standards for educational and psychological testing* (AERA, APA, & NCME, 1999) requires that when test score equating procedures are used to produce comparable scores, detailed technical information about the equating method and data collection method used should be provided, and indices measuring the uncertainty in the estimated equating function should also be estimated and reported. In practice, the accuracy of equating is commonly assessed through standard errors of equating (von Davier, Holland, & Thayer, 2004; Kolen & Brennan, 2004) that reflect the degree of sampling errors in equating functions. Equating biases are also indices used to assess the quality of equating.

2.3 The Equipercentile Equating Function in the Observe-Score Equating

The derivation of the equipercentile equating function is detailed in this section because it is used by many observed score equating such as the PSE and the CE methods. Test score equating methods can be divided into two different categories: the observed score equating and the item response theory (IRT) equating. This study focuses on the observed score equating method. The most important component of observed score equating methods is the equipercentile equating function (von Davier, Holland, & Thayer, 2004).

The equipercentile equating function is developed by identifying scores on the new form (Y) that have the same percentile ranks as scores on the old form (X). For

example, to find a form Y equivalent of a Form X score, one has to start by finding the percentile rank of the Form X score. Then it has to do with finding the Form Y score that has the same percentile rank. Algebraically, the equipercentile equating function for converting Y scores to X scores, $e_X(y)$, is obtained by (Kolen & Brennan, 2004)

$$e_X(y) = F^{-1}[G(y)],$$

where x and y are, respectively, a particular value of X and Y, F^{-1} is the inverse of the cumulative distribution function F, G is cumulative distribution function of Y.

2.4 The Non-Equivalent Groups with Anchor Test Design (NEAT)

A wide range of test equating designs can be used for collecting equating data but the non-equivalent groups with anchor test (NEAT) design (von Davier, Holland, & Thayer (2004) which is also called common-items non-equivalent groups design (Kolen Brennan, 2004) is the most frequently used in practice. It is because of the fact that when groups are nonequivalent, some information is needed to adjust for group differences and typically that information is scores on the anchor test. In the NEAT design (see Table 1), the two operational tests to be equated, X and Y, are given to two samples of examinees from potentially different test populations (referred to as P and Q). In addition, an anchor test, A, is given to both samples from P and Q.

Samples from P and Q that take the test at different administrations are generally self selected and thus might differ in systematic ways. One of the well-known systematic differences between P and Q are ability differences. Adjustments are needed to compensate for such differences using the appropriate anchor test which can be either internal or external test. The internal anchor test is a part of X and Y, while the external

anchor test is used only to adjust for group differences but it is not used for scoring the test. It is recommended that the anchor test should be proportionally representative of the two tests in content and statistical characteristics.

When groups differ greatly, all equating methods tend to produce large equating errors because the anchor test fails to adjust for group differences. In practice, the degree of the precision of equating results could be assessed through estimates of standard errors of equating (SEE).

Table 1. Non-Equivalent groups with Anchor Test Design (NEAT)

Population	Sample	*X*	*A*	*Y*
P	1	✓	✓	Not observed
Q	2	Not observed	✓	✓

It is suggested that to enhance the equating performance of the anchor test in producing more accurate equating results, the construction of the anchor tests should be specially created so that their correlations with the tests to be equated are maximized (Sinharay and Holland (2006, 2007). More specifically, the anchor test can be constructed by embedding items with moderate difficulty. Doing so relaxes the requirement of equal distributions of statistical characteristics between the anchor tests and the operational tests. This newly suggested anchor test construction provides an alternative guideline for constructing the anchor test as it is an approach for enhancing the precision of the test equating function.

2.5 The Importance of Synthetic Population

The designed missing data as parts of the NEAT design makes available a variety of equating methods that use the NEAT design. Because X is never observed for examinees in Q and Y is never observed for examinee in P (see Table 1), different types of missing data treatments are required to handle these missing data. For these reasons, there are several different methods of test score equating under the NEAT design (Holland, & Dorans, 2006) such as the Tucker method, and the Levine method. However, two commonly used observed score equating methods that use the NEAT design are the chain equipercentile (CE) method and the post-stratification equating (PSE) method which is also known as the frequency estimation method (Kolen & Brennan, 2004). These two methods are different in the way that the anchor test score is used to produce an equating function and also in different assumptions about missing data made to handle missing data arising when the NEAT design is used to collect equating data (Holland, Sinharay, von Davier, & Han, 2007).

Unlike the CE, the PSE method is considered a strong equating method because it is more sound in terms of the developed theoretical foundation (Haris, & Kolen, 1990; Kolen, 1992), making this equating method more appealing than the CE. Braun and Holland (1982) noted that there are theoretical problems with the CE which center on the definition of the equipercentile equating function. That is, equipercentile relationships are defined for a particular group of examinees, as in the PSE method. However, the equipercentile relationship between X and Y for the CE method is not defined for a particular group. More specifically, the PSE method uses synthetic population functions defined as the weighted mixture of P and Q as an important tool to equate test scores,

while the CE method does not define any synthetic population function (Kolen, 1992). The reason for employing synthetic population functions was detailed by Braun and Holland (1982). The basic reason of this development is that the NEAT design uses "two samples of P and Q" which sometimes are nonequivalent groups—they are different to some degrees depending on for example how well they are sampled. However, an equating function for the NEAT design is typically viewed as being defined for a single population. To obtain a "single" population for defining a single equating relationship, therefore P and Q must be combined (Kolen & Brennan, 2004). Given the appeal of the synthetic population function, this study focuses on the PSE method.

To obtain a single population for defining an equating relationship, Braun and Holland (1982) used the target population (T) or the synthetic population function which is the weighted mixture of P and Q to combine P and Q. T is an important ingredient of the PSE method for performing equating. T is a mixture of both P and Q, $T = wP + (1-w)Q$, where w is the weight given to P. The weight can be any number raging from 0 to 1 but in Holland, & Rubin (1982) the usual w is a proportion of sample size from P relative to the total sample size (P+Q) defined by $w = N_P/(N_P + N_Q)$, where N_P and N_Q represent the sample sizes of P and Q, respectively.

2.6 Presmoothing the Score Distribution

In computing equipercentile equating functions, the estimates of population score distributions can be used in place of the observed sample distributions. The estimated score distributions are typically much smoother than the distributions observed in the sample (Livingston, 1993). Therefore, the estimated distributions are often described as

smoothed, and the process is often referred to as *smoothing*. The PSE method uses smoothed distributions in computing equating relationship.

Log-linear models (Holland & Thayer, 2000) are a smoothing model that offers the user a flexible choice in the number of parameters to be estimated from the data. The log-linear smoothing model is detailed as follows.

Assume there is a random variable X that defines the test form X with possible values $x_0,...,x_j$, with j is the possible score values, and the corresponding vector of observed score frequencies $n=(n_0,...,n_j)^t$ that sum to the total sample size N. The vector of the population score probabilities $p=(p_0,...,p_j)^t$ is said to satisfy a log-linear model if

$$\log_e(p_j) = \alpha + u_j + b_j\beta$$

Where the (p_j) are assumed to be positive and sum to one, b_j is a row vector of constants referred to as score functions, $\boldsymbol{\beta}$ is a vector of free parameters, u_j is a known constant that specifies the distribution of the (p_j) when $\boldsymbol{\beta} = 0$, and α is a normalized constant that ensures that the probabilities sum to one.

When u_j is set to zero, the log-linear model used to fit a univariate distribution is

$$\log_e(p_j) = \alpha + \sum_{i=1}^{I} \beta_i (x_j)^i.$$

The terms in this model can be defined as follows: the $(x_j)^i$ are score functions of the possible score values of test X (e.g., $x_j^1, x_j^2, ..., x_j^I$) and the β_i are free parameters to be estimated in the model fitting process. The value of I determines the number of moments of actual test score distribution that are preserved in the smoothed distribution. For example, if I=4 then the smoothed distribution preserves the first, second, third, and fourth moments (mean, variance, skewness, and kurtosis) of the observed distribution.

Similarly, the bivariate distribution of the scores of two tests (e.g., X and Y) is given by

$$\log_e(p_{jk}) = \alpha + \sum_{i=1}^{I} \beta_{xi}(x_j)_i + \sum_{h=1}^{H} \beta_{yh}(y_k)_h + \sum_{g=1}^{G} \sum_{f=1}^{F} \beta_{gf}(x_j)^g (y_k)^f ,$$

Where p_{jk} is the joint score probability of the score (x_j, y_k; score x_j on test X and score y_k on test Y). This model produces a smoothed bivariate distribution that preserves I moments in the marginal (univariate) distribution of X; H moments in the marginal (univariate) distribution of Y; and a number of cross moments ($G \leq I$, $F \leq H$) in the bivariate X-Y distributions.

2.7 Post-Stratification Equating (PSE) Method

The process of equating test score using the PSE method is composed of two major steps. The first step is to estimate score frequencies of missing data by invoking conditional assumptions such that score frequencies of missing data are obtained for constructing synthetic population functions. Anchor test scores are usually used as the conditional variable for the PSE method (von Davier, Holland, & Thayer, 2004). So in

this study, the traditional PSE method is referred to the PSE method that uses anchor test scores. The second step is to use derived synthetic population functions to equate test score on the new form (Y) to score on the old form (X) using the equipercentile equating function. Equated scores are scores on the new form (Y) that have the same percentile ranks as scores on the old form (X).

In order to create the T for X and Y (T_X and T_Y), score distributions from both populations must be known but, as seen in Table 1, scores on X for the population Q and scores on Y for the population P are however unavailable due to the characteristic of the NEAT design (known as designed missing). Therefore some statistical assumptions need to be invoked to obtain the score distributions for the missing parts to be used for constructing the synthetic population.

Test equating methods that use the NEAT design employ different untestable statistical assumptions about the uses of anchor test data to predict the scores on the designed missing parts. The post-stratification equating (PSE) method assumes that the conditional distributions of X conditional on the anchor test data (A) are the same across populations and it is similar for the distributions of Y conditional on A, which are expressed by

$$f(x \mid A,P) = f(x \mid A,Q) \qquad (1)$$

and

$$f(y \mid A,Q) = f(y \mid A,P) \qquad (2)$$

Let f_{xP} be the marginal distribution of X for the population P, f_{xQ} the marginal distribution of X for the population Q, f_{yQ} the marginal distribution of Y for the

population Q, and f_{yP} the marginal distribution of Y for population P, then the distributions for the synthetic population for Form X and Y are

$$f_{(x)} = w_X f_{xP} + (1 - w_X) f_{xQ} \tag{3}$$

$$f_{(y)} = (1 - w_Y) f_{yP} + w_Y f_{yQ}. \tag{4}$$

The quantities f_{xQ} in *(3)* and f_{yP} in (4) are usually missing data (unobserved) but can be obtained as follows by using assumptions (1) and (2):

$$f_{xQ} = \sum_a f(x, A = a \mid Q) = \sum_a f(x \mid A = a, P) h_{aQ} \tag{5}$$

$$f_{yP} = \sum_a f(y, A = a \mid P) = \sum_a f(y \mid A = a, Q) h_{aP} \tag{6}$$

where h_{aQ} and h_{aP} are marginal distributions of A for the population Q and P, respectively. The expression in (5) and (6) can be substituted into (3) and (4), correspondingly, to provide expressions for the synthetic population as follows:

$$f_{(x)} = w_X f_{xP} + (1 - w_X) \sum_a f(x \mid A = a, P) h_{aQ} \tag{7}$$

$$f_{(y)} = (1 - w_Y) \sum_a f(y \mid A = a, Q) h_{aP} + w_Y f_{yQ} \tag{8}$$

Then equating scores on X to scores on Y based on the synthetic population functions can be carried out using the equipercentile equating function mentioned earlier. This equating function is analogous to the equipercentile relationship for random groups equipercentile equating function (Kolen & Brennan, 2004).

Even though the PSE is theoretically a promising method, under general realistic conditions, the PSE equating relationship does not correspond to the relationship for the

CE method (Braun & Holland, 1982). Moreover, it produces larger equating biases than does the CE method (e.g., Wang, Lee, Brennan, & Kolen, 2008; Holland, von Davier, Sinhary, & Han, 2008). The reason for this shortcoming might be because the PSE method employs the less reasonable missing data assumption when compared to the CE method (Holland, Sinharay, von Davier, & Han, 2008) which does not require any assumption about missing data. It is later found that there is more evidence revealing that using scores on the anchor test to make the conditional assumption as a way to deal with the missing data assumption of the PSE method is less reasonable. By using anchor test true scores to replace anchor test scores, the result of the PSE method is however much improved and more accurate than that of the CE (Wang & Brennan, 2009). This method is called the modified frequency estimation method.

Therefore, it comes to understand that when attempting to improve the equating result of the PSE method it is better to use a good variable as a conditional variable to predict frequencies of missing data in the NEAT design and anchor test true score is a promising choice. Another attempt proposed by Holland and Sinharay (2007) is to construct an anchor test with items with medium difficulty, but their method is appropriate for an external anchor test only. However, the use of anchor test true score may not be sufficient to remove biases when P and Q differ greatly.

This study proposes to use subscores combined with anchor test scores in two different ways to handle missing data and group differences. The two methods will be explained in the next sections. Using both subscores and anchor test scores in this study is based on the idea that using more information to predict score frequencies of missing data is expected to increase the accuracy of equating results. By using the combination of

subscores and anchor test score, the prediction of score frequencies of missing data could be improved because high correlations between operational test scores and subscores could have a potential to increase the accuracy of the prediction of missing data. This method is also viable when a number of good examinee demographic variables are not available to compute the propensity score (Rosenbaum & Rubin, 1982) which is a recommended conditional variable to be used to handle group differences (Livingston, Dorans, & Wright, 1992). Although using examinees' demographic variables combined into examinees' propensity score to adjust for group differences is recommended, a large number of demographic variables is recommended because a smaller set of demographic variables used to compute propensity scores could not add much value to the equating results (Paek, Liu, &Oh, 2008)

2.8. Subscore Estimation

There has been much interest in assisting students in determining which of the skills within a particular domain of knowledge needs improvement and numerous testing programs report subscale scores defined by the test design. Most of achievement tests have subsections and a subscore is the score assigned to a subsection of the test. Subscore reporting usually provides more detailed diagnostic information about examinees' performance that may be useful, for example, in making individual instruction placement and remediation decisions (Tate, 2004) and in formatively supporting teaching and learning (Dibello & Stout, 2007).

Tests with multiple subsections imply a mulidimensional structure of tests. Using scores on subtests may provide additional information about examinee performance rather than using only total test scores. To estimate and report subscores of a test,

sophisticated approaches such as multidimensional item response theory (MIRT) can be used. Alternatively, subscores can also be estimated using the classical test theory (CTT) where subscale scores are estimated from number-correct responses. The Haberman method (2008) of subscore estimation, which is based on CTT, was adopted in this study because it produced estimates of subscores that that were highly correlated with estimates from the MIRT approach (Haberman, & Sinharay, 2008). Moreover, it is straightforward and does not require much computation time.

The methodological approach to subscore estimation is illustrated and detailed in Haberman (2008) and Sinharay, Haberman, and Puhan (2007). The Haberman method of subscore estimation is typically a regression of true subscore on both observed score and observed total score, and the linear regression of true subscore τ_X on the observed subscore S_X and the observed total score S_Z is estimated by

$$L(\tau_X \mid S_X, S_Z) = E(S_X) + \beta(\tau_X \mid S_X \cdot S_Z)[S_X - E(S_X)] + \beta(\tau_X \mid S_Z \cdot S_X)[S_Z - E(S_Z)],$$

where

$$\beta(\tau_X \mid S_X \cdot S_Z) = \frac{\sigma(\tau_X)[\rho(S_X, \tau_X) - \rho(\tau_X, S_Z)\rho(S_X, S_Z)]}{\sigma(S_X)[1 - \rho^2(S_X, S_Z)]}$$

and

$$\beta(\tau_X \mid S_Z \cdot S_X) = \frac{\sigma(\tau_X)[\rho(S_Z, \tau_X) - \rho(\tau_X, S_X)\rho(S_X, S_Z)]}{\sigma(S_Z)[1 - \rho^2(S_X, S_Z)]}$$

This method of true subscore estimation gives weights to both the total score and the subscore and provides a better approximation of a true subscore than is provided by observed subscore alone (Haberman, 2008).

2.9. Propensity Score Method

Faced with the potential selection bias resulting from nonequivalent groups, researchers performing observational studies have become increasingly interested in statistical adjustments to the estimates of treatment effects based on the propensity score (Rosenbaum & Rubin, 2006).

Propensity scores (Rosenbaum & Rubin, 2006) are the estimated conditional probability that a subject will be assigned to a particular treatment, given a vector of observed covariates. It is considered a one-dimensional summary of multidimensional covariates such that when the propensity scores are balanced across the two groups, the distributions of all covariates, X, are balanced in expectation and across the two groups (D'Agostino & Rubin, 2006.). In other words, propensity score analysis is the process by which the attempt is to balance nonequivalent groups by estimating each participant's conditional probability of treatment assignment using observed covariates. Then one can use these probabilities for case matching, stratification, covariate adjustment, or weighting of observation.

When brought to equating context, treatment assignment can be regarded as "test form assignment" and the observed covariates are examinees' demographic variables. Even though examinees take different test forms, if they are homogenous in terms of the propensity score, their distribution of covariates is the same, no matter what test form is administered to them. It is therefore reasonable to assume that any groups of examinees who have the same propensity scores would have identical distributions of their total scores, which would be the realistic assumptions for the PSE method. This argument is similar to the assumption made by Wang and Brennan (2009). The only difference is that

this study uses the propensity score as the conditional variable instead of the anchor test true score used in Wang and Brennan.

In the observational study, propensity scores are computed from examinees' information such as collateral information (Mislevy & Sheehan, 1989). Once the examinee's information is obtained, they are combined into the examinee's propensity score that represents an examinee's likelihood of being assigned to a particular test form (e.g., test form 1). Using the propensity score estimation is a way to achieve this goal and estimated propensity scores can be used for case matching, stratification, covariate adjustment, or weighting of observation. In practice, the examinee information commonly used to compute propensity scores are examinees' demographic variables. Ideally, the demographic variable that is appropriate for estimating propensity score should be a variable that can distinguish the two groups of examinees.

This study used subscores, the anchor test score, and demographic variables to compute examinees' propensity scores. This set of examinees' information is called collateral information about examinees in this study. Once examinees' collateral information is obtained, they are combined into examinees' propensity scores using a statistical modeling. The propensity score has a value ranging from 0 to 1.00. Any examinees taking different test forms and having equivalent propensity scores would be balanced in collateral information.

Numerous propensity score methods such as logistic regression, classification trees, bootstrap aggregation, and boosted regression have been proposed in the literature in the observational study to estimate propensity scores (Luellen, 2007). Even though the equating literature has not illustrated yet what method is the most effective in producing

propensity scores, it was found in the observational study that the logistic regression method worked well at reducing bias and tended to result in more precise estimates of treatment effect with less potential for introducing bias (Luellen, 2007). Therefore, this study used the logistic regression method to compute examinees' propensity scores.

Logistic regression is a form of statistical modeling that is often appropriate for binary outcome variables (that is, data y_i that take on the values 0 or 1). It describes the relationship between a binary outcome variable and a set of covariates. A logistic regression has applications in various fields such as medicine and social science research and its advantages is that the model interpretation is possible through odds ratios, which are functions of model parameter. The logistic regression function is given by

$$P(y=1 \mid x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)},$$

where $P(y=1)$ denotes probability of Y=1, representing its dependence on values of explanatory variables (X), α denotes the intercept and β the coefficient. When the model holds with β=0, the binary response is independent of X. The simplest way to interpret a logistic regression coefficient is in terms of "odds ratios." If two outcomes have the probabilities (p, $1-p$), then $p/(1-p)$ is called the odds. The log odds of logistic function has a linear relationship (Agresti, 1990, p. 86) which is given by

$$\log\left(\frac{\Pr(y=1 \mid x)}{\Pr(y=0 \mid x)}\right) = \alpha + \beta x \ .$$

2.10 Multiple Imputation Method

Missing data often occurs due to factors beyond the control of the researcher. Missing data may be planed. For example, they are part of the research design which is

similar to missing data arising due to the NEAT design. Missing data can create biases in parameter estimates that can lead to generalization problems. When missing data is serious, valid inferences regarding a population of interest cannot be made. This occurs when missing data is not missing at random. For example, it is missing in a manner which makes the sample different from the population from which it was drawn.

There are several methods developed to handle missing data such as listwise deletion, pairwise deletion, and imputation of missing data (replace the missing data with estimated scores). The multiple imputation method (Rubin, 1987) has increasingly gained interested to researchers in various fields because it has been shown to produce unbiased parameter estimates (Schafer & Graham, 2002). In the multiple imputation method, missing values for any variable are predicted using existing values from other variables (covariates). The predicted values are called "imputes", and are substituted for the missing values, resulting in a full data set called an "imputed data set." This process is performed multiple times producing multiple imputed data sets (hence the term "multiple imputation"). The results from *m* imputed data sets are analyzed using standard statistical analyses and the results from *m* complete data sets are combined to produce inferential results. It is recommended that small number of *m* (e.g., 5 imputations) is adequate for multiple imputation (Fichman and Cummings, 2003) but larger is better when fraction of missing data is large (Schafer, 1997).

Currently, multiple imputation procedures are more accessible to researchers. One can impute missing data using software such as SAS. The SAS V9.1software "(SAS Institute, 2003) has a procedure "*proc mi* that enables one to impute missing data easily.

2.10.1 Missing data mechanism

Data are missing for many reasons. For example, participants dropped out from a longitudinal study, died, or refused to answer surveys. In some cases, missing data is a result of a research design itself. The data collection that uses the NEAT design is an example of a design that creates missing data.

Missing data are problematic because most statistical procedures require a value for each variable. When a data set is incomplete, the data analyst has to decide how to deal with it. Causes of missing data fit into three categories, which are based on the relationship between the missing data mechanism and missing and observed values. The first is missing completely at random (MCAR). MCAR means that the missing data mechanism is unrelated to the values of any variables. The second is missing at random (MAR). MAR means that the missing values are related to either observed covariates or response variables. When missing data is MCAR or MAR, the missing mechanism is ignorable and the best method to use to impute data is the multiple imputation method that uses maximum likelihood (ML). The third is not missing at random (NMAR). NMAR means that missing values depend on missing values themselves. When missing data is NMAR, the missing data mechanism is non-ignorable.

When equating data are collected using the NEAT design, examinees taking X will have missing values on Y, and examinee taking Y will have missing values on X. Missing data of the NEAT design is said to be missing by design. Conventionally, it has been assumed that this missing data mechanism is MAR (Holland & Rubin, 1982). That is, score distributions of two different groups of examinees are assumed to be identical when the anchor test score is held constant. When missing data mechanism is MAR, the

anchor test score can be used to fill in missing values. This assumption is feasible when there are no group differences in terms of ages or abilities. However, when groups differ greatly in abilities, this assumption is likely to fail. In the literature, certain background information has been recommend for matching observations when the anchor test score fails to reduce equating biases due to group differences (Livingston, Dorans, & Wright, & Dorans, 1992).

2.10.2 Introduction to the EM algorithm

The Expectation Maximization (EM) algorithm is a very general iterative algorithm for ML estimation in complete-data problems. Analyses performed using the EM algorithm assumes that missing data are MAR (Little & Rubin, 2002). Basically, the EM algorithm employs iterative processes in which initial estimates of missing data values are obtained. Basically, the EM algorithm employs these steps: (1) replace missing values by using estimated values, (2) estimate parameters, (3) re-estimate the missing values, assuming the new parameter estimates are correct, (4) re-estimate parameters, and so fourth, iterating until convergences.

Each iteration of EM consists of an E step (expectation step) and an M step (maximization step). Each step has a direct statistical interpretation. Specifically, the E step finds the conditional expectation of the missing data given the observe data and current estimated parameters, and then substitutes these expectations for the missing data. The M step performs ML estimation of parameters just as if there were no missing data, that is, as if they have been filled (Little and Rubin, 2002).

One of advantages of the EM algorithm is that it can be shown to converge reliably. However, when there is a large fraction of missing information, its rate of convergence can be painfully slow.

2.11 The Statement of Research Problem and its Solution

It has been shown that the anchor test fails to remove equating biases when groups differ greatly in abilities. To reduce equating biases produced by the PSE method, the anchor test score should be replaced with the anchor test true score (Wang & Brennan, 2009). Although the anchor test true score is an interesting choice in that it can reduce equating biases of the PSE method, it is not clear if it would produce the accurate equating result when samples of P and Q differ greatly in abilities. Group differences may arise due to some reasons. For example, two samples of P and Q taking the test at different administrations are "generally self selected and thus might differ in systematic ways" (Peterson, Kolen, and Hoover, 1993). One of the well-known systematic differences between the P and Q are the ability differences and therefore adjustments are needed to compensate for such differences by using the appropriate anchor test. It was found that when groups differ greatly in ability, all equating methods that use the NEAT design produce larger equating biases and standard errors of equating because the distributions of the scores on the anchor test in the two groups are not the same (von Davier, 2003). Larger equating biases and standard errors of equating occur when the anchor test score does not adjust for group differences well (Holland & Sinharay, 2007), implying that the conditional assumptions about missing data do not hold in this case. These findings imply that using only anchor test true score may produce less equating accuracies when groups differ greatly in ability. The possible reason is that using a single

piece of examinee information (e.g., anchor test score) to adjust for group differences may not be achieved.

In order to obtain more accurate equating results when groups differ greatly, it is suggested that multiple pieces of examinee information such as their demographic variables be used together with scores on the anchor test to make the conditional distribution assumptions more appropriate. Such information may be combined into a form of the propensity score (Rosenbaum & Rubin, 1983), where the examinee' s propensity score is a conditional probability that the examinee will be assigned to a particular test form, given a vector of observed covariates. Paek, Liu, and Oh (2008) found that using a small number of demographic variables did not add much value to improve equating results. It has been suggested that it would be useful to choose variables that distinguish the two groups of examinees to establish examinees' propensity scores (Livingston, Dorans, & Wright, 1990). The variables that are of interest include variables that are related to opportunity-to-learn (Kolen, 1990).

This study proposed the use of subscores combined with the anchor test score and demographic variables to improve equating results of the PSE method. Subscores are increasingly attractive to researchers, educators, and policy makers for the diagnostic purposes, but their premises have not been explored in the equating context. More accurate equating results might be obtained by using subscores because high subscore to total score correlations can produce more accurate score frequencies of missing data needed for constructing synthetic population functions and deriving equating functions. The use of more collateral information about examinees defined as available information

about examinees in additional to test scores is an alternative to the traditional PSE method that uses the anchor test score only.

Once collateral information is ready, it can be used to compute score frequencies of missing data due to the design of the NEAT in two different ways. The first method combines them into a form of examinees' propensity scores with which the anchor test score is replaced to compute such score frequencies. The second method is to use observed collateral information to impute missing data using the existing multiple imputation method. This study examined if these two different implementations improve the PSE equating results.

2.12 The Goal of This Study and the Evaluation Indices

The goal of this study is to evaluate the effectiveness of using collateral information about examinees for the PSE method in two different aspects: the accuracy of the prediction of score frequencies of missing data; and the accuracy of equating in terms of equating biases, and standard errors of equating. This investigation was explored using both simulation data and the empirical data. The following details research questions of this study.

The first question addressed in this study is related to whether the additional use of the collateral information by the purposed methods (the propensity score method and the multiple imputation method) offer the improvement to the synthetic population functions. Specifically, to assess whether it adds the additional improvement to synthetic population functions is actually to assess the improvement of the prediction of score frequencies of missing data predicted by purposed methods. This can be assessed by evaluating the agreement between observed score frequencies of full data (pseudo-test

data) and predicted score frequencies of missing data. This study adopted the agreement indices of Holland et al (2008), where these indices include Pearson χ^2 statistic, and Likelihood ratio χ^2 statistic.

The second research question of this study is related to the performance of the proposed methods of this study in terms of accuracy of equating functions. The investigation is carried out by evaluating the equating biased and standard errors of equating. The criterion equating function used for computing the equating bias is the PSE equating function obtained by using the pseudo-test data or full data which is generated data without missing data.

The third research question concerns the relative performance of the proposed methods in establishing accurate equating functions. Given that the anchor test score and the anchor test true score have been used as conditional variables to make missing data assumptions of the traditional PSE method and the modified PSE method, respectively, this investigation is carried out to evaluate the equating performance of the proposed methods by comparing equating accuracy indices (the equating biases, and standard errors of equating) produced by the proposed methods and the other methods. The method that produces the smallest equating biases and standard errors of equating is the most appropriate for test score equating. The criterion equating function used for computing the equating bias is the PSE equating function obtained by equating test scores using the pseudo-test data.

CHAPTER III

RESEARCH METHOD

The data analysis of this study has two parts: simulation and empirical data analyses. The simulation data were used to evaluate if it was feasible to use collateral information about examinees to improve accuracy of the post-stratification equating (PSE) method in terms of equating biases, and standards errors of equating. There were three simulation factors investigated in this study including ability differences, test length, and missing data treatment. For real data analyses, group differences and missing data treatment were investigated.

This chapter details the research design, data generation, and the equating procedure for the simulation study. For empirical data analysis, descriptive statistics and the information about the empirical data set chosen for this study are presented. Note that the data analyses for the empirical data were the same as those for the simulation data.

3.1 Research Design

This study examined the feasibility of using collateral information about examinees (sub-scores, anchor test score, and examinees' demographic variables) in two different ways (the propensity score stratification method and the multiple imputation method) as an effort to improve the accuracy of the PSE method. Simulation factors were manipulated as follows.

3.1.1 Test Length

This study investigated two different test lengths, 60 and 40 items, which represent long and short tests respectively. Details regarding how tests were generated are presented in the next section.

3.1.2. Ability Differences between Groups of Examinees

Previous studies found that group differences had tremendous effects on both standard errors of equating and equating biases. This study investigated two types of ability differences between two groups of examinees. The first type was the condition in which there were no ability differences between two groups of examinees. The second type was the condition in which the two groups differ greatly in abilities. More specifically, the group of examinees taking the second test form or the new test form (the *Q* population) was a group that was more proficient than the group taking the old form (the *P* population). How to manipulate the degree of group differences is presented in the next section.

3.1.3 Missing data treatment

Two methods of missing data treatment used in this study included the propensity score stratification method and the multiple imputation method. When collateral information was obtained, they were manipulated in eight different sets to investigate which collateral information set yielded best equating results for the PSE method. As noted previously, this study proposed using collateral information about examinees in the PSE method to compute score frequencies of missing data. These frequencies are needed for equating test scores using the PSE method. It was interesting to examine what types of collateral information provided better improvement to equating accuracy. Therefore, for each of 8 conditions, the following different sets of collateral information were investigated.

- Anchor test score only (A), which is the traditional PSE method
- Anchor test true score only (T), which is the modified PSE method

- Demographic variables only (D)
- Anchor test and demographic variables (A&D)
- Anchor test score and sub-scores (S&A)
- Sub-scores and demographic variables (S&D)
- Anchor test score, sub-scores and demographic variables (ALL)

This yielded 64 conditions (8*8=64) investigated in this study, and 100 data sets were replicated for each condition for the propensity score method. But 20 data sets were replicated for the multiple imputation method.

3.2 Data Simulation Procedure

This study equated scores on two test forms (*X* and *Y*). Tests with multiple subsections imply a multidimensional structure. It was reasonable in this study to use a compensatory multidimensional item response theory (MIRT) model (e.g., Reckase, 1997) to generate *X* and *Y* scores. To generate examinees' multiple test scores or sub-scores which are sums of correct responses within each subsection, the items parameters had to be generated. Then item responses for the five subtests were generated using the compensatory MIRT model. The procedure for the item parameter generation was explained in the next section.

The probability of a correct response to item *i* of examinee *j* for the compensatory MIRT model can be expressed as

$$P(X_i = 1 \mid \theta) = c_i + (1 - c_i)\frac{\exp[1.7a_i'\theta_j - b_i]}{1 + \exp[1.7a_i'\theta_j - b_i]},$$

where

X_i is the score (0, 1) on item i (i=1, …, n),

a_i' is the vector of item discrimination parameters (slope),

b_i is the scalar difficulty parameter for item i,

c_i is the scalar guessing parameter for item i, and

θ_j' is the vector of trait parameters for person j (j=1, …, N).

3.2.1 Item Parameter Generation

It was assumed in this study that each of the two test forms to be equated had five content areas (subsections). Item responses of two test forms (X and Y) were simulated using a 5-dimensional IRT model. Note that there were 60 items for the long test condition and 40 items for the short test condition. For the long test condition, 45 items were operational test items (each section had 15 items), and 15 items were the anchor test items. For the short test condition, 30 items (each section had 6 items) were operational test items, and 10 items were anchor test items. Item parameters were generated separately for the long test and short test condition.

Specifically, item parameters for the 5-dimensional IRT model were generated as follows. The vector of slope ($\boldsymbol{a}$), difficulty (b), and guessing (c) parameters for the compensatory MIRT model were generated using WINGEN2 (Han & Hambleton, 2007) such that $\boldsymbol{a}$~LN(0, .2), b~N(0,1), and c~BETA(8,32), where LN(μ, σ) designates a log-normal distribution with mean μ and standard deviation σ of the logarithm, and BETA(α, β) a beta distribution with two parameters α and β. Note that Form X and Form Y item parameters were generated using the same procedure. That is, two sets of item parameters (for X and Y) were generate separately.

Item parameters produced by the WINGEN2 were a complex structure, meaning that items, approximately, measure multiple dimensions equally. This was inconsistent with item specifications in real practices; items are usually purposely developed to measure only one dimension. Therefore, modifications were made to produce a more simple structure. This was done by allowing items to be dominantly loaded on a single dimension only, by fixing ***a***-parameters corresponding to other dimensions at .01. For example, in the long test condition, items 1 to 12 had higher *a*-parameters on the first dimension than dimensions 2 to 5, by replacing ***a***-parameters for dimensions 2 to 5 with .01. Similarly, items 13 to 24 had higher *a*-parameters on the second dimension than other dimensions, by replacing *a*-parameters for dimension 1, and ***a***-parameters for dimensions 3 to 5 with .01.

After item parameters were generated, operational tests (*X* and *Y*) and the anchor test (A) were constructed as follows. For the long test condition, the first 9 items from each of the five subsections were chosen as operational test items. Therefore, there were 45 items chosen for *X* and 45 items chosen for *Y*. The last 3 items from each of the five subsections were chosen as anchor test items. Therefore, there were 15 anchor test items. For the short test condition, the first 6 items from each of the five subsections were chosen as operational test items. Therefore, there were 30 items chosen for *X* and 30 items chosen for *Y*. The last 2 items from each of the five subsections were chosen as anchor test items. Therefore, there were 10 anchor test items.

To item parameters for an anchor test, the Form *Y* generating common item parameters were replaced with the Form *X* generating common item parameters. This procedure was used for both long and short test conditions.

3.2.2 θ Parameter Generation

This study did not simulate examinees' demographic variables because simulating demographic variables was infeasible due to the fact that the distribution of demographic variables and abilities was unknown. Therefore, demographic variables from the real data were used and merged with the simulated test data, pretending that simulated examinees had those demographic variables. However, by doing so, the sample sizes in this study could not be varied and were fixed at 1,361 and 1,266 for test form 1 and 2, respectively.

The five vector of theta estimates were obtained by calibrating empirical data using a multidimensional IRT model. Specifically, a test form 1 and a test form 2 were separately fitted to a 5-dimensional item response theory model using the software WinBUGS1.4 (Spiegelhalter, Thomas, Best, & Lunn, 2003) to obtain examinees' five ability estimates. The WinBUGS code used to run WinBUGS is in the appendix C—this code was modified from the code written by Bolt and Lall (2003). The correlation coefficients among the estimates of θ parameters for the test forms 1 and 2 are in Table 2 and Table 3, respectively. These correlation matrices are roughly similar, indicating that covariance structures for the test form 1 and 2 data are roughly comparable.

Table 2. Correlation coefficients among θ parameters from WinBUGS (test form 1)

θ_1	1.00				
θ_2	.09	1.00			
θ_3	.55	.15	1.00		
θ_4	.16	.10	.12	1.00	
θ_5	.01	.31	.14	.54	1.00

Table 3. Correlation coefficients among θ parameters from WinBUGS (test form 2)

θ_1	1.00				
θ_2	.21	1.00			
θ_3	.37	.03	1.00		
θ_4	.05	.24	.27	1.00	
θ_5	.19	.01	.10	.47	1.00

As presented in Table 4, averages of the five ability estimates for each test form were close to zero as the result of the parameterization in WinBUGS that sets the means of estimates to zeros. Five vectors of theta estimates for the test form 1 and five vectors of theta estimates for the test form 2 were used along with the generated item parameters mentioned above to generate item responses for the condition that there were no group differences. However, when there were group differences in terms of abilities, 0.5 was added to the five vectors of ability estimates for the test form 2, meaning that examinees taking the test form 2 were more proficient in all five dimensions than those examinees taking the test form 1. The differences in the mean abilities of .5 were used because small differences cannot allow for investigating values of collateral information to the improvement of equating biases. Therefore, greater differences about .4 or .5 in standard deviation unit of θ were used (e.g., Holland and Sinharay, 2007).

Table 4. Average θ parameters from WinBUGS

	Average				
	θ_1	θ_2	θ_3	θ_4	θ_5
Test form 1	0.00	0.00	0.01	0.00	0.00
Test form 2	-0.01	0.01	-0.01	0.01	0.02

3.2.3 Item Responses Generation

The simulation data were generated using the SAS V9.1 software (SAS Institute, 2003). Specifically, the probability of a correct response to item *i* by simulated examinee *j* was computed using the compensatory multidimensional item response theory model (e.g., Reckase, 1997). A response vector of dichotomous item scores for each examinee was obtained by generating, for each item, a uniform random number (ranging between 0 and 1) and comparing the value with the probability of an examinee of that ability level passing the item. If the computed probability exceeded the random number, then the item was scored as correct (1); otherwise, it was scored as incorrect (0).

This study did not generate examinees' demographic variables, but used demographic variables from the empirical data set. After item responses data were generated, they were merged with the demographic variables from real data. Specifically, the examinees' demographic variables were linked up with the examinees' simulated test scores by matching examinees' demographic variables with their item responses using the estimates of their θ values from WinBUGS. Therefore, every simulation data set has the same demographic values, and the sample sizes were not varied. Note that test form *X*

and test form *Y* data generated were called "pseudo-test data" and they had no missing data.

To check if data was acceptably generated, one generated data set was analyzed through an exploratory factor analysis and a confirmatory factor analysis using the Mplus5.2 (Muthen & Muthen, 2008), which is a structural equation model software. It was found that the exploratory factor analysis produced a 5-factor model that was slightly better than the 6-factor model. For example, the BIC was 100,083.801 for the 5-factor model, while it was 100,370.907 for the 6-factor model. Moreover, the confirmatory factor analysis of the simple structure of this test showed that the 5-factor model fitted data very well as indicated by the small and non-significant chi-square statistic (χ^2 =692.484, df = 670, p=.2658). This evidence indicated that the 5-dimension data was reasonably generated and therefore it was reasonable to use the data generation procedure to generate test scores used for equating in this study.

3.3 Missing Data Generation

There were two types of simulation data used in this study. The first type of data was the complete test data, and the second type was the incomplete test data. Each of these data was used to address different research questions. Data used to address the research questions 1 to 2 are the pseudo-test data (Holland, Sinharay, von Davier, & Han, 2007), which is the complete test data. The pseudo-test is the test that is manipulated by pretending that each examinee has scores on both test forms, meaning that it was pretended that each examinee took both test form *X* and *Y*. The creation of this type of data was proposed and used to test the assumptions of various equating methods that use the NEAT design (Holland, Sinharay, von Davier, & Han, 2007). Later, it was used in the

TEDS-M project to investigate the effectiveness of the balanced incomplete block design used to collect the international assessment data of the TEDS-M project.

3.3.1 Pseudo Test Data and Missing Data Generation for 60-Item Test

Under the MIRT model, the Form-X and Form-Y generating item parameters were used, respectively, to generate item responses X for P and item responses Y for P, resulting in a number of examinees of P having 105 item responses (45 items for test form X, 45 items for test form Y, and 15 anchor test items). Similarly, Form-X and Form-Y generating item parameters were used, respectively, to generate item responses X for Q and item responses Y for Q, resulting in a number of examinees of Q having 105 item responses (45 items for test form X, 45 items for test form Y, and 15 anchor test items). Data for P and Q were then merged and called complete data. A completed data set was called a pseudo-test which was used as the criterion for comparisons. For example, observed frequencies of score X for Q and observed frequencies of score Y for P were used as the true frequencies. Also, the criterion equating functions used for computing biases were obtained by equating scores on pseudo-tests (completed data).

An incomplete test data set that reflects missing data from the NEAT design was created from the pseudo-tests data simulated above, by deleting the first 45 item responses X from Q and the last 45 item responses Y from P. But the 15 anchor test items were kept in both forms.

3.3.2 Pseudo Test Data and Missing Data Generation for 40-Item Test

For the 40-item test conditions, pseudo test data and test data with missing data were generated by the same procedure as for the 60-item test conditions. Specifically, the generating item parameters for the test form X and test form Y were used to generate item

responses X for P and item responses Y for P. The same sets of generating item parameters were used to generate item responses X for Q and Y for Q. P and Q were then merged, pretending that each examinee had a score X and score Y. This procedure resulted in examinees having 70 item responses (30 items for the test form X, 30 items for the test form Y, and 10 anchor test items).

To generate test data from the NEAT design, the first 30 item responses X were deleted from Q and the last 30 item responses Y were deleted from P. But the remaining 10 anchor test items were kept in both forms.

The mean ($\overline{X}$), standard deviation (S.D.), minimum (Min.), and maximum (Max.) for the simulated test form X and Y data are presented in Table 5 and Table 6, respectively. As shown in these Tables, when there were group differences, the mean for the test form 2 is greater than that for the test form 1. For example, the mean for the 60 item-test form 2 is 41.66 (S.D. = 2.05), which is greater than that for the test form 1 (mean = 35.62, S.D. =2.00). However, when there are no group differences, scores on test form 2 and test form 1 are identically distributed.

Table 5. Descriptive statistics for the simulated test form 1

		$\overline{X}$	S.D.	Min.	Max.
No Ability Differences	60 items	35.62	2.01	30.47	41.34
	40 items	24.66	1.45	20.52	28.91
Ability Differences	60 items	35.62	2.00	29.57	41.53
	40 items	24.66	1.43	20.75	29.14

Table 6. Descriptive statistics for the simulated test form 2

		$\bar{X}$	SD.	Min.	Max.
No Ability Differences	60 items	36.99	2.05	29.89	44.00
	40 items	23. 31	1.40	18.40	28.64
Ability Differences	60 items	41.66	2.05	34.50	48.25
	40 items	26.48	1.44	21.77	31.01

3.4 Analytic Strategies for the Purposed Equating Methods

This study had two parts. The first part was to evaluate the predictions missing data from the combination of sub-scores and the anchor test score. The second part evaluated equating results obtained by using the two methods proposed in this study. Therefore, this research method section comprised of two sections corresponding to these two parts. Specifically, the first section explained the procedure to predict missing data and how to assess the prediction performance of collateral information, while the second section described procedures and how to equate test scores using the proposed two methods.

3.4.1 Prediction of Score Frequencies of Missing Data

There were two proposed uses of sub-scores to equate test scores. The first method was to use the examinees' propensity score as a stratification variable in the equating process of the PSE method by replacing the anchor test score with the propensity score. The second method uses a multiple imputation methods

to compute missing scores directly. When applied to equate test scores, these two proposed methods used different strategies to estimate score frequencies of missing data required for constructing the synthetic population functions.

3.4.1.1 The Propensity Score Approach to the Prediction of Missing Data

Holland, Sinharay, von Davier, and Han (2007) provided the approach to predicting score frequencies of missing data of the NEAT design and their strategies were adopted in this study. The procedure to predict score frequencies of missing data using the propensity score method is as follows.

First, a logistic models was used to estimate the propensity score (Z) of each examinee by using collateral information as predictor variables. The detailed sub-scores estimation and propensity score estimation are presented in Section 3.4.2.1 and Section 3.4.2.3, respectively. Then the loglinear model was used to presmooth the bivariate distribution of (X, Z) obtained from P and the bivariate distribution of (Y, Z) from Q, by preserving the first four moment of X and Y and covariance between Y and Z; and X and Z. The presmoothed bivariated probabilites, respectively, are denoted as:

$$p_{xz} = \mathrm{P}\{X=x, \mathrm{Z}=\mathrm{z}|P\} \text{ and } q_{yz} = \mathrm{P}\{Y=y, \mathrm{Z}=\mathrm{z}|Q\}$$

These bivariate probabilities were used to form the marginal distributions of Z in P and Q, that is

$$h_{zP} = \sum_{x} p_{xz} \text{ and } h_{zQ} = \sum_{y} q_{yz}$$

Then the conditional probability, $\mathrm{P}\{X=\mathrm{x}|\mathrm{Z}=\mathrm{z}, P\}$, was computed as the ratio p_{xz}/h_{zP}. The estimated conditional probabilities were used to obtain the predicted score probabilities for X in Q as follows:

$$f_{xQ} = \sum_z p_{xz}(h_{zQ} / h_{zP}).$$

By similar reasoning, the predicted score probabilities for Y in Q are

$$f_{yP} = \sum_z p_{yz}(h_{zP} / h_{zQ})$$

The predicted frequencies of X in P and Y in Q, respectively, were $N_Q f_{xQ}$ and $N_P f_{yP}$, where N_Q and N_P were sample sizes of Q and P respectively. The focus of this section was to assess the agreement of $N_Q f_{xQ}$, the observed frequencies of X in Q (n_{xQ}), the agreement of $N_P f_{yP}$ and the observed frequencies of Y in P (m_{yP}). The criteria used for this investigation include Pearson chi-square statistic (χ^2), Likelihood ratio chi-square statistic (G^2), and Freeman-Tukey (FT) residuals. The following formulas define these statistics. In each case, n_i denotes the observed frequencies and m_i the corresponding predicted frequencies:

Pearson chi-square statistic, $\chi^2 = \sum_i \frac{(n_i - m_i)^2}{m_i}$,

Likelihood ratio chi-square statistic, $G^2 = 2\sum_i n_i \log(n_i / m_i)$,

Freeman-Tukey (FT) residuals, FT residuals = $\sqrt{n_i} + \sqrt{n_i +} - \sqrt{4m_i + 1}$

These three statistics are often used to measure the closeness of the fitted frequencies to observed frequencies in discrete distributions of score (Holland & Thayer, 2000). Note that χ^2 and G^2 measure a summary of the closeness between the observed frequencies and predicted frequencies. However, FT residuals assess the closeness at each score

point. FT residuals are also used to assess the rounding effect of the multiple imputation method at each score point.

3.4.1.2 The Multiple Imputation Approach to the Prediction of Missing Data

When the NEAT design is used to collect equating data, the missing data occurs because of the unique characteristic of the NEAT design that creates the designed missing. That is, scores on test form X for Q and scores on test form Y for P are never observed. The multiple imputation method using "*proc mi*" in the SAS V9.1 software (SAS Institute, 2003) was used in this study to impute these missing data. The procedure "mi" implemented in the SAS V9.1 was appropriate for the NEAT design because missing data mechanism generated by the NEAT design is assumed to be missing at random (MAR; Rubin, 1976) or an ignorable mechanism. In other words, the subpopulation the two groups represent are assumed to have the same target score distribution when the anchor test score is held constant (Holland & Rubin, 1982). The EM algorithm in SAS V9.1 assumes that a missing data mechanism is MAR. When missing data mechanism is ignorable, the EM algorithm is appropriate to impute missing data.

In this study, the EM algorithm implemented through the procedure *mi* in SAS V9.1 (SAS Institute, 2003) was used to compute test score X for the population Q and test score Y for the population P. Specifically, this study used 20 simulation data sets each having 5 imputation data sets. Five imputations were chosen because five imputations are considered to be adequate in the multiple imputation (Rubin, 1996; Schafer, 1997; Fichman & Cummings, 2003). Then imputed values were rounded. Any imputed values less than 0 were set to 0, and any values greater than the maximum score point was set to

the maximum value of score points. The effect of rounding was trivial at the low and high ends of the score scale as seen in the result section.

For the imputation *ith* (i=1,..., 5) of the data set *jth* (j=1,..., 20), once X for Q and Y for P were predicted, the predicted score frequency distribution of X for Q and predicted score frequency distribution of Y for P were obtained directly, which are denoted by f_{xQ} and f_{yP}, respectively.

Similarly to Section 3.4.1.1, the chi-square statistic, likelihood ratio chi-square statistic, and FT residuals were computed to assess the agreement between the predicted score frequencies and the observed (true) score frequencies. These statistics were averaged across 5 imputations and 20 data sets and the resulting averages were used and reported.

3.4.2. Test Score Equating Procedure

This study focused on the PSE method, and the procedure for equating test scores using methods of this study employed the following steps. These steps were based on the procedures of the PSE equating method (von Davier, Holland, & Thayer, 2004; Kolen & Brennan, 2004) but little modifications were made such that the propensity scores were included in the PSE equating framework. The modified steps were as follows:

1. Estimate subscores using CTT model
2. Place estimates of subscores on the same scale
3. Estimate propensity score using the logistic regression model
4. Construct synthetic population functions
5. Equate test score using the equipercentile method

However, when the multiple imputation method was used, step 3 was not necessary and thus it was replaced by the multiple imputation approach to score frequency estimation. Also, step 4 is less complicated than when the multiple imputation method was used. The following section has more details regarding the procedure of equating test scores.

3.4.2.1 Sub-score Estimation

This study generated tests with five subsections. The sub-score estimation used in this study is based on the classical test theory (CTT). The sub-score estimation used was adopted from Haberman (2008) and Sinharay, Haberman, and Puhan (2007). The Haberman method of subscore estimation is a regression of true subscore on both observed subscore and observed total score, and the linear regression of true subscore τ_X on the observed subscore S_X and the observed total score S_Z is estimated by

$$L(\tau_X \mid S_X, S_Z) = E(S_X) + \beta(\tau_X \mid S_X \cdot S_Z)[S_X - E(S_X)] + \beta(\tau_X \mid S_Z \cdot S_X)[S_Z - E(S_Z)],$$

where

$$\beta(\tau_X \mid S_X \cdot S_Z) = \frac{\sigma(\tau_X)[\rho(S_X, \tau_X) - \rho(\tau_X, S_Z)\rho(S_X, S_Z)]}{\sigma(S_X)[1 - \rho^2(S_X, S_Z)]}$$

and

$$\beta(\tau_X \mid S_Z \cdot S_X) = \frac{\sigma(\tau_X)[\rho(S_Z, \tau_X) - \rho(\tau_X, S_X)\rho(S_X, S_Z)]}{\sigma(S_Z)[1 - \rho^2(S_X, S_Z)]}$$

This method of true sub-score estimation gives weights to both the total score and the sub-score and provides a better approximation of true sub-score than is provided by observed sub-score alone (Haberman, 2008).

3.4.2.2. Place the Estimated Sub-scores on the Same Scale

The estimates of true sub-scores of different test forms may have different scales and different meaning especially when two groups of examinees differ greatly in abilities. Therefore, it is necessary to adjust the estimated sub-scores by using information that is common across two groups of examinees. In this study, the estimates of true sub-score were adjusted by the covariate adjustment technique, where the covariate used was the anchor test score.

Specifically, the j^{th} sub-score for the i^{th} examinee $L(\tau_X \mid S_X, S_Z)_{ij}$ was adjusted as follows:

$$La_{ij} = L(\tau_X \mid S_X, S_Z)_{ij} - \beta_j(A_i - \overline{A}),$$

where β_j is the regression weight of a sub-score L on anchor test score (A), $\overline{A}$ denotes the mean of the anchor test score, and A_i the score on the anchor test of the examinee i.

3.4.2.3. Estimate Propensity Scores

Examinees' propensity scores (Z) were estimated using the logistic regression model. The outcome variable was the test form (F) (F=0 if test Form=X, otherwise 1), and a set of covariates includes five sub-scores, demographic variables, and the anchor test score. An estimated examinee's propensity score is the predicted group membership of an examinee assigned to the test Form Y. The estimated examinees' propensity scores were then divided into 21 strata.

3.4.2.4. Construct synthetic population

Synthetic population (T) is a mixture of both P and Q, $T = wP + (1-w)Q$, where w is the weight given to P. When the propensity scores (Z) were used to construct synthetic population, the two assumptions of the post-stratification equating (PSE) method (von Davier, Holland, & Thayer, 2004; Holland, & Dorans, 2006) were modified as follows:

1. The conditional distribution of X given Z over T, $f(X=x|\ Z=z,\ T)$, is the same for any T of the form $T=wP+(1-w)Q$.
2. The conditional distribution of Y given Z over T, $f(Y=y|\ Z=z,\ T)$, is the same for any T of the form $T=wP+(1-w)Q$.

By using the assumptions 1 and 2, the score distributions of X and Y for T were estimated by

$$f_{(x)_T} = w_X f_{xP} + (1-w_X)\sum_z f(x \mid Z=z,P)h_{zQ}$$

$$f_{(y)_T} = (1-w_Y)\sum_z f(y \mid Z=z,Q)h_{zP} + w_Y f_{yQ}$$

where $h_{zQ} = P(Z=z \mid Q)$ and $h_{zP} = P(Z=z \mid P)$

However, when the multiple imputation method was used to estimate missing data, the propensity score (Z) was not used because the multiple imputation method is capable of using the collateral information to compute missing data for P and Q directly. In addition, PSE assumptions mentioned above were not necessary for the multiple imputation method synthetic population functions were constructed directly from the imputed score X and imputed score Y. Specifically, the procedure "*mi*" implemented in SAS was used to impute missing data. The EM algorithm was used with the procedure

"*mi*" because missing at random is assumed for the equating data obtained through the NEAT design. "*Proc mi*" imputed examinees' total scores *X* or *Y*, while the conditional variables were the collateral information about examinees. When missing test scores *X* and *Y* data were filled, the score distributions of *X* and *Y* for the target population *T* were then estimated by

$$f_{(x)_T} = w_X f_{xP} + (1 - w_X) f_{xQ}$$

and

$$f_{(y)_T} = (1 - w_Y) f_{yX} + w_Y f_{yQ}.$$

Note that $f(x)$ and $f(y)$ were smoothed distributions obtained by presmoothing *X* and *Y* separately using the log-linear model (Holland & Thayer, 2000). The first five moments of score distributions were preserved to eliminate irregular distributions of imputed values.

3.4.2.5. Equate Test Scores

Note that $f(x)$ and $f(y)$ are the distributions of test score of *X* and *Y*, respectively. The equating that transforms *X*-raw score to *Y*-raw score was carried out using the equipercentile function which is

$$\text{Equi}_Y(x) = G^{-1}(F(x)),$$

where *G* is the cumulative distribution of *Y*, *F* the cumulative distribution function of *X*, and G^{-1} the inverse of cumulative distribution function *G*.

3.4.2.6. Evaluation Criteria

For the propensity score method, the criteria for evaluating equating results included equating bias, and standard errors of equating (SE). The following formulas define these measures:

$$\text{Bias(x)} = \frac{1}{100}\sum_{i=1}^{100}[\hat{e}_i(x) - e(x)],$$

where $\hat{e}_i(x)$ is the equated score for x at each replication , $e(x)$ is the criterion equating functions obtained from the equating of scores of the pseudo-test data (without missing data).

The standard error of equating for a score point x was estimated by

$$\text{SE(x)} = \left(\frac{1}{100}\sum_{i=1}^{100}[\hat{e}_i(x) - \bar{\hat{e}}(x)]^2\right)^{\frac{1}{2}}$$

where $\bar{\hat{e}}(x)$ is the mean across 100 replications The method that produces the least bias and SE is the one that works best for equating test scores.

However, for the multiple imputation method, only 20 simulation data sets were used. However, 5 data sets were imputed for each of the 20 data set. Therefore, standard errors of equating and equating biases were computed differently from the propensity score method. The standard errors of equating and equating biases for the multiple imputation method were computed as follows.

For the data set j^{th} (j=1,..., 20) that has 5 imputed data sets (i=1, ..., 5), the standard error of equating for a score point x for the data set j^{th} was computed by

$$\text{SE(x)}_j = \left(\frac{1}{5} \sum_{i=1}^{5} [\hat{e}_i(x) - \bar{\hat{e}}(x)]^2 \right)^{\frac{1}{2}},$$

where $\hat{e}_i(x)$ is the equated score for x at each replication, $\bar{\hat{e}}(x)$ is the mean across 5 replications. Since there were 20 replications, the standard errors of equating reported in this study is the average SE (x) across the 20 replications, which was computed by

$$\text{SE(x)} = \frac{\sum_{j=1}^{20} SE(x)_j}{20}.$$

The equating biases for the data set i^{th} were computed as

$$\text{Bias(x)}_j = \frac{1}{5} \sum_{i=1}^{5} [\hat{e}_i(x) - e(x)]$$

where $e(x)$ is the criterion. Similarly, the equating bias for a score point x reported in this study is the average equating biases across the 20 replications, which was computed by

$$\text{Bias(x)} = \frac{\sum_{j=1}^{20} Bias(x)_j}{20}.$$

3.5. Real data Analysis

Real data analyses were performed to address the research questions 3 and 4. The research questions 1 and 2 were not investigated using real data in this study because data in which examinees have scores on both forms were not available. Therefore, only

standard errors of equating and equating biases were compared. The empirical data is from a cross-national study of the preparation of middle school mathematics teachers called MT21 (Schmidt, et al, 2007).The comparisons were based on the responses of students (future mathematics teachers) from the six participating countries (A, B, C, D, E, and F). Data were collected from teachers in their first or last year of preparation by sampling institutions in each country. Futures teachers were questioned on their (1) background, (2) course taking and other program activities, (3) knowledge relevant to teaching—mathematical and pedagogical, and (4) beliefs and perspectives on content and pedagogy.

The equating data are future teachers' test scores on the mathematics content knowledge. There are two different test forms, each of which measures five content area—algebra, data & interpretation, function, geometry, and number. The distribution of MT21 items across the five content areas is in Table 7. Table 8 shows test performance of the six participating countries.

Table 7. Distribution of Items of MT21 Data

content	Form			Total
	Form1	Form2	Anchor	
Algebra	2	12	8	22
Data	4	3	5	12
Function	10	3	6	19
Geometry	8	7	2	17
Number	14	5	3	22
Total	38	30	24	92

Table 8. Test score Performance of Six Countries

	$\overline{X}$		S.D.		Min.		Max.	
Country	Form 1	Form 2	Form 1	Form 2	Form 1	Form 2	Form 1	Form 2
A	24.09	20.77	10.41	9.76	0	4	45	45
B	29.06	27.48	10.56	8.65	0	0	55	47
C	41.82	35.08	5.97	4.43	21	16	53	44
D	22.66	22.61	6.01	5.57	1	2	39	36
E	41.38	36.89	5.81	4.95	21	19	53	50
F	27.29	27.86	7.71	6.05	6	12	55	45

MT21 data were manipulated to investigate the impact of group differences on standard errors of equating and equating biases. That is, there were two conditions investigated: no group differences and group differences. For the no group differences condition, all cases in the MT21 data were used. The descriptive statistics of the empirical data are summarized in the Table 9. Table 10 shows correlations between operational test scores and anchor test scores as well as correlations between operational test scores and sub-scores.

Table 9. Descriptive Statistics of the Empirical data

Form	Test	n	$\overline{X}$	SD	Min	Max	Reliability
Form 1	X	1361	20.22	6.94	0	35	.83
	Anchor	1361	10.48	4.70	0	21	.81
Form2	Y	1266	17.34	4.68	0	27	.75
	Anchor	1266	10.85	4.56	0	21	.80

Table 10. Correlations Between the Operational Test Score and Sub-scores

	Form	Anchor	Algebra	Data	Function	Geometry	Number
Score	X	.65	.38	.57	.68	.69	.76
	Y	.61	.62	.48	.43	.73	.53

However, when there were group differences, country E and country F data were excluded from the test form 1, while country A and country D data were excluded from the test form 2. This resulted in that examinees taking the test form 2 were more proficient than examinees taking the test form 1. Table 11 shows descriptive statistics for this condition and Table 12 shows correlations between operational test scores and anchor test scores as well as correlations between operational test scores and sub-scores.

Table 11. Descriptive Statistics of Empirical Data (Group Differences Condition)

Form	Test	n	$\overline{X}$	SD	Min	Max	Reliability
Form 1	*X*	925	17.37	6.51	0	34	.86
	Anchor	925	9.66	4.45	0	23	.76
Form2	*Y*	1009	18.41	4.11	0	27	.84
	Anchor	1009	13.08	4.97	0	23	.80

Table 12. Correlations Between the Operational Test Score and Sub-scores (Group Differences Condition)

	Form	Anchor	Algebra	Data	Function	Geometry	Number
Score	X	.52	.35	.66	.78	.64	.74
	Y	.55	.72	.44	.47	.74	.55

For the propensity score method, analyses of real data were carried out using the Kernel equating software because it provides a convenient way to estimate standard errors of equating. However, for the multiple imputation method, 5 data sets of test data (X and Y) were imputed and each was equated using the SAS macro. Standard errors of equating were computed based on the results of the five imputed data sets using the same equation as mentioned in the simulation data analysis section

Equating biases were obtained by comparing a resulting equating function to an equating function obtained from IRT true score equating. An IRT true score equating function was obtained by fitting both test form X and test form Y simultaneously to a 3-parameter logistic IRT model using the software BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2002). For each test form, the estimates of items parameters and the examinee's theta value were used to compute the examinee's probabilities of responding to MT21 items correctly. A sum of these probabilities was the examinee's true score. Then true scores of examinees taking X and Y were equated using the KE software (Chen, Yan, Hemat, Han, & von Davier, 2008) and the resulting equating function was used as the criterion to compute equating biases.

CHAPTER IV

RESULTS

The main objective of this study was to explore if it is feasible to use collateral information about examinees in the post-stratification equating (PSE) method in order to improve the quality of equating in terms of standard errors of equating and equating biases. Collateral information used in this study included sub-scores, anchor test scores, and demographic variables. The traditional PSE equating method and the modified PSE method use the anchor test true score and the anchor test true score, respectively, to adjust for group differences. When groups differ greatly in ability, it is necessary to use more collateral information to adjust for the differences so as to reduce biases due to unintended differences that cannot be eliminated by the anchor test score. It was hypothesized in this study that using more information to construct synthetic functions used to equate test scores under the PSE method could reduce equating biases. In other words, since constructing synthetic population functions deals with predicting score frequencies of missing data. When equating data are missing due to the data collection design called "Non-Equivalent Groups with Anchor Test" (NEAT) design, using more collateral information to impute score frequencies of the missing data was expected to gain "predicted score frequencies" that are more close to "true frequencies." Therefore, it was expected that collateral information can reduce equating biases. However, this study also explored the impact of collateral information on standard errors of equating. In addition, the simulation part of this study assessed how close "score frequencies imputed by using collateral information" and "true frequencies" were using chi-square statistics, likelihood ratio chi-square statistics, and Freeman-Tukey (FT) residuals.

This chapter presents equating results (standard errors of equating and equating biases) of equating methods that used collateral information about examinees in the post-stratification equating method, in comparison with the traditional PSE method and the modified PSE method. Collateral information was used in two missing data treatments (the multiple imputation method and the propensity score post-stratification method) to obtain score frequencies of missing data used to equate test score using the PSE method.

This study investigated the impact of ability differences between groups of examinees and test length on standard errors of equating and equating biases. Since collateral information was expected to reduce biases due to group differences, ability differences between groups of examinees were manipulated and examined. Test length was included as another factor because sub-scores have more values for a long test than for a short test. In practice, both short and long tests are used. Therefore, it was necessary to assess the impact of test lengths on equating accuracies, when collateral information was used as the stratification variable in the PSE method. The sample size is a factor worth studying, but this study did not include the sample size as a simulation factor because it was challenging to simulate examinees' demographic variables, and demographic variables that have potentials to reduce equating biases are not well documented in equating literature.

Simulation study enables the factors mentioned above to be investigated. Simulation factors in this study included two treatments of missing data (propensity score stratification method and multiple imputation method), two ability differences between two groups of examinees (no differences vs. group differences) and two test lengths (long test condition [60 items—45 items plus 15 anchor test items] and short test condition [40

items—30 items plus 10 anchor test items]). The combination of these factors yields eight simulation combinations. For each of the eight conditions, different equating strategies listed below were conducted to investigate different effects of collateral information on standard errors of equating and equating bias.

1. Anchor test sore only (A)
2. Anchor test score, sub-scores, and demographic variables (ALL)
3. Anchor test score and demographic variables (A&D)
4. Sub-scores and demographic variables (S&D)
5. Demographic variables only (D)
6. Sub-scores only (S)
7. Anchor test true score (T).
8. Sub-scores and anchor test (S&A)

Therefore, there were 64 (8*8) conditions investigated. For the propensity score stratification method of missing data treatment, 100 simulated data were replicated. Equating biases and standard errors of equating were computed based on the 100 replications. For the multiple imputation (MI) method, the first 20 data sets of the simulated 100 data sets were used. Five data sets were imputed for each of the chosen 20 data sets, resulting in 100 analyses. The equating biases and standard errors of equating were summarized across the 20 imputed data sets.

There are two main sections presented in this chapter. The first section presents results of equating of simulation data. The equating results that were the target of considerations include standard errors of equating, equating biases, Pearson chi-square statistics, likelihood ratio (LR) chi-square statistics, and FT residuals. Standard errors of

equating and equating biases were used to assess accuracies of equating. Pearson chi-square statistics, LR statistics, and FT residuals were used to evaluate how well each of the eight sets of collateral information recovered the true score frequencies. Better recovery of score frequencies of missing data is associated with smaller equating biases.

The second section presents the equating results using empirical data. To explore what sets of collateral information about examinees were more effective in terms of standard errors of equating and equating biases, empirical data were manipulated such that differences in ability between groups of examinees were varied. Specifically, when there were no group differences in abilities, all cases in the empirical data set were used. But when there were group differences, the country *C* and country *E* data were excluded from the test form 1, and the country A and country D data were excluded from the test form 2.

4.1 The Results from Simulation Data

4.1.1 The Propensity Score Method: Standard Errors of Equating

When the propensity score method was used to obtain score frequencies of missing data of the NEAT design to construct synthetic population functions (X and Y) as required for equating test scores using the PSE method, standard errors of equating and equating biases were computed from 100 replicated analyses. Figures 4.1a to 4.1d present standard errors of equating. Specifically, Figure 4.1a presents standard errors of equating in the long test condition when there were no group differences. Figure 4.1b presents standard errors of equating in the longer test condition when group differences were greater. Figure 4.1c presents standard errors of equating in the short test condition when

there were no group differences. Figure 4.1d presents standard errors of equating in the short test condition when group differences were greater.

One objective of this study was to compare PSE methods that use collateral information to the traditional PSE method and the modified PSE method in terms of standard errors of equating. For the longer test condition when there were no group differences, Figure 4.1a shows that all equating methods had large standard errors of equating. Three equating methods that had smaller standard errors of equating included the traditional PSE method which is the PSE method that uses the anchor test score (A), the modified PSE method (T), and the PSE method that uses the combination of the anchor test score and demographic variables (A&D). Standard errors of equating of the PSE method that uses demographic variable (D) were larger; the values ranged from 3.7 to 9.0. The standard errors of equating for the PSE method that uses the combination of sub-scores and demographic variables (S&D) were larger at the low end, medium in the middle, and smaller at the high end of the score scale. When all collateral information about examinees (ALL) was used in the PSE method, standard errors of equating were large.

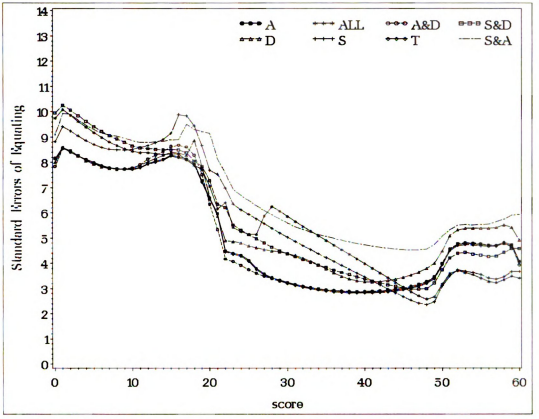

FIGURE 4.1a. PS standard errors of equating: long test and no group differences

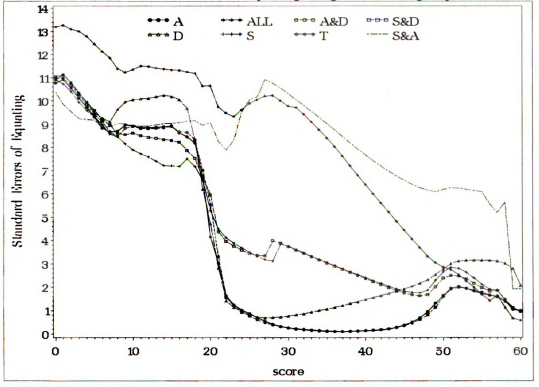

FIGURE 4.1b. PS standard errors of equating: long test and group differences

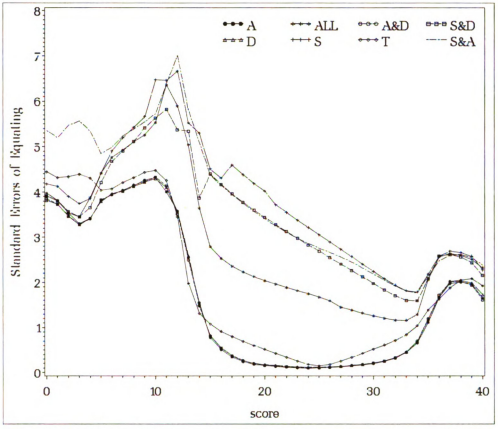

FIGURE 4.1c. PS standard errors of equating: short test and no group differences

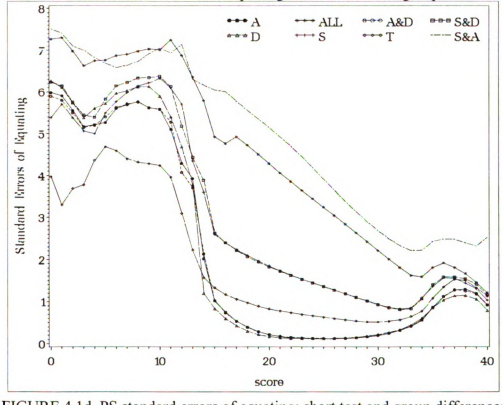

FIGURE 4.1d. PS standard errors of equating: short test and group differences

When group differences in terms of abilities were greater (Figure 4.1b), all methods had standard errors of equating more different from each other. The traditional PSE method (A) and modified PSE method (T) had the smallest standard errors of equating. The PSE methods that used demographic variables (D), the method that use the combination of sub-scores and demographic variables (S&D), and the method that uses sub-scores (S) had larger standard errors of equating approximately within a range from 2.0 and 10.0. The PSE method that uses all collateral information about examinees (ALL) and the method that uses the combination of subs-scores and anchor test score (S&A) had the largest standard errors of equating.

For the short tests condition, Figures 4.1c and 4.1d shows that standard errors of equating were similar between when there were no group differences and when there were group differences. Specifically, four equating methods that had the smallest standard errors of equating included the traditional PSE method (A), the PSE method that uses demographic variables (D), the PSE method that uses the combination of the anchor test score and demographic variables (A&D) and the modified PSE method (T). The methods that involved sub-scores had large standard errors of equating. For examples, the PSE method that uses all collateral information about examinees (ALL), the PSE methods that used sub-scores (S), the PSE method that use the combination of sub-scores and demographic variables (S&D), and the PSE method that uses the combination of sub-scores and the anchor test score (S&A) had the largest standard errors of equating at the middle of the scale score. Standard errors of equating of all PSE equating methods were poorly estimated at the low end and high end of the score scale, and the explanation for this finding is discusses in chapter V.

4.1.2 The Propensity Score Method: Equating Bias

As noted in Chapter III, one objective of this study was to compare PSE methods that use collateral information to the traditional PSE method and the modified PSE method in terms of equating biases. The comparisons are presented in Figures 4.2a to 4.2d. For the longer test condition (60 items), equating biases produced by different equating methods were comparable, except for at the lower end of the score scale. Larger equating biases at the low end of score scale might be associated with zero frequencies. Even though all equating methods had comparable equating biases, the method that uses demographic variables (D) had smaller equating biases than the traditional PSE method (A) and the modified PSE method (T). When groups differed greatly in abilities, Figure 4.2c shows that all methods produced even more different equating biases. Specifically, they had larger positive biases. The positive impact of sub-scores on equating biases was evident when groups differ in ability in the long test condition. That is, all methods that used sub-scores outperformed other methods. Specifically, the PSE method that used sub-scores (S), the PSE method that used sub-scores and demographic variables (S&D), and the PSE method that used sub-scores and anchor test (S&A) were the three methods that had smaller equating biases at the middle of the score scale (20 to 50) , even though equating biases they produced were large. But their equating biases at the low end of the score scale were large. The method that uses demographic variables (D) performed best at the low end and high end of the score scale. The traditional PSE method (A) and the modified PSE method (T) had comparable equating biases; they had larger equating biases than the methods that involve sub-scores at the middle of the score scale, but smaller at the low and high ends of the score scale.

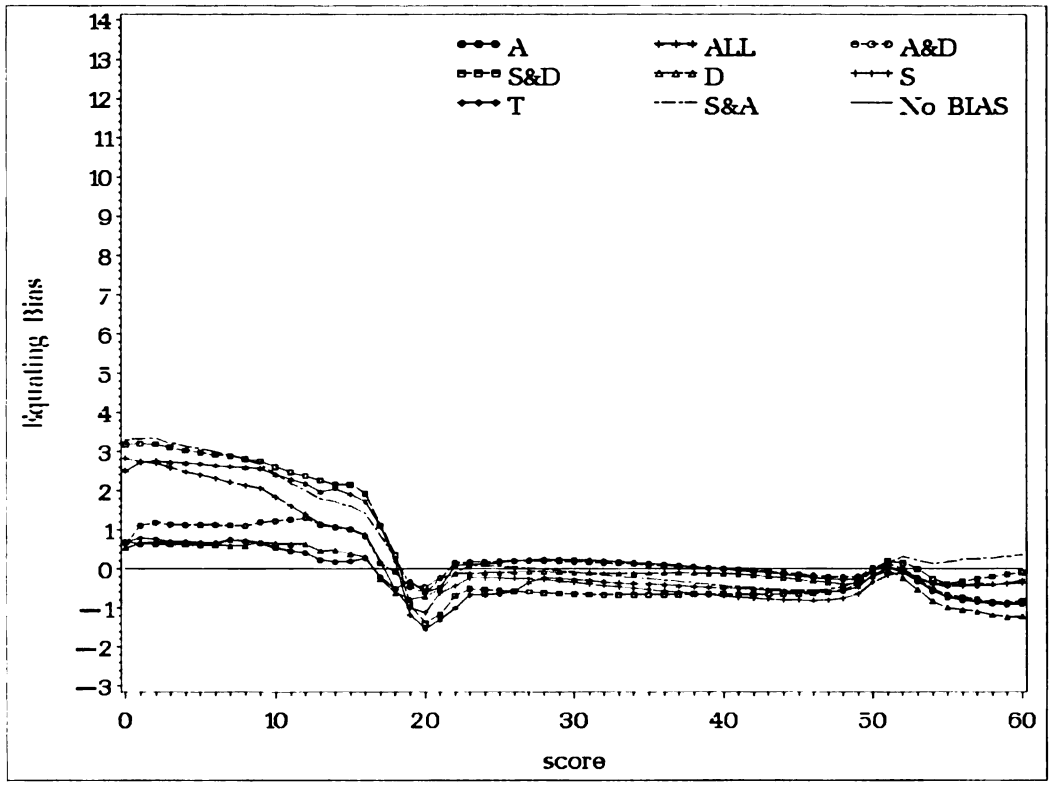

FIGURE 4.2a. PS equating biases: long test and no group differences

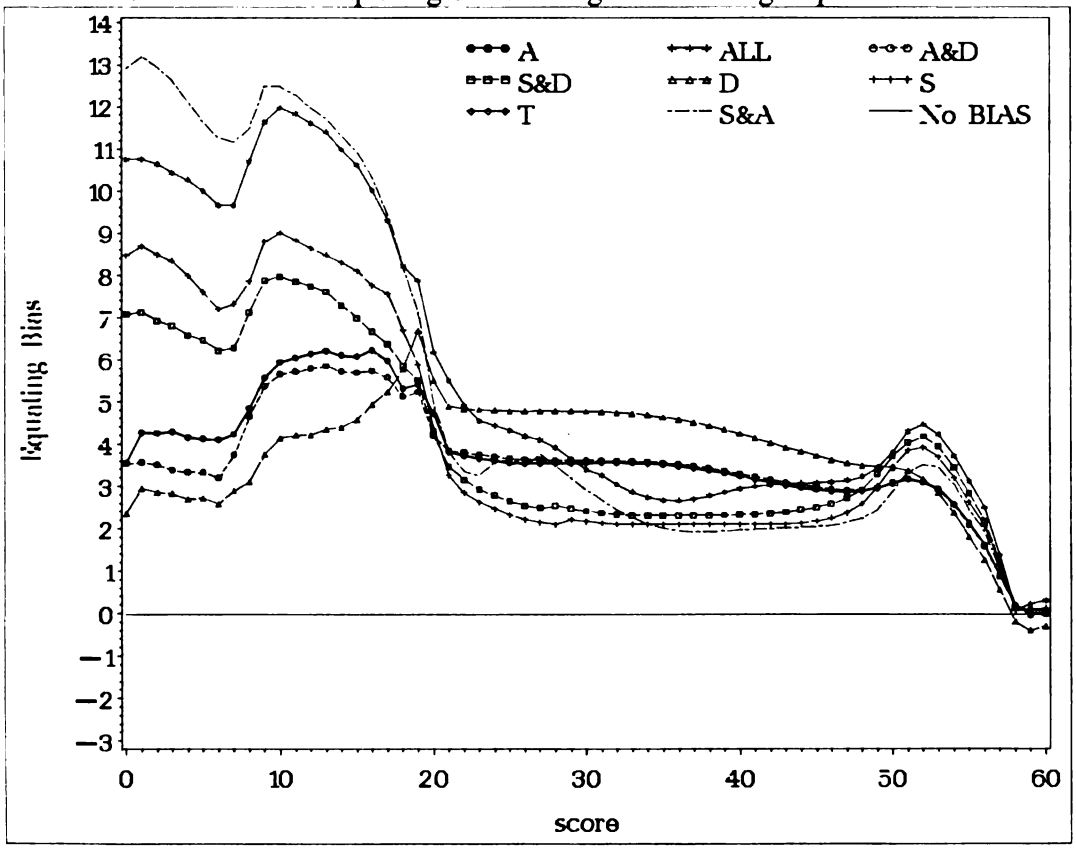

FIGURE 4.2b. PS equating biases: long test and group differences

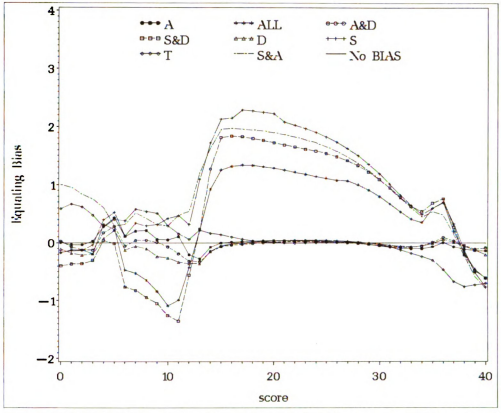

FIGURE 4.2c. PS equating biases: short test and no group differences

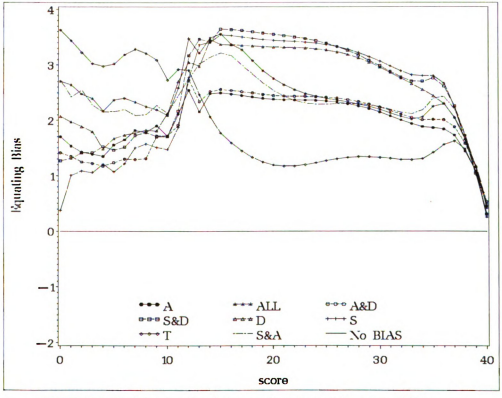

FIGURE4.2d. PS equating biases: short test and group differences

For the short tests condition (40 items), Figure 4.2c shows that the traditional PSE method (A), the modified PSE method (T), and the method that uses the combination of the anchor test score and demographic variables (A&D) had the smallest and comparable equating biases at the middle of the score scale. Unlike in the longer test condition, using sub-scores in the short test condition resulted in larger equating biases. For example, as seen in Figure 4.2c, the methods that used sub-scores (S, S&A, S&D, ALL) all produced larger positive equating biases than other methods. Such large positive equating biases were more evident at the middle of the score scale, which were about 1.5 to 2.0 points larger than other methods. When groups differ in abilities, all methods produced positive equating biases as shown in Figure 4.2d. As noted earlier, it was evident that the methods that used sub-scores performed the least. The modified PSE method (T) had the smallest equating biases at the score scale of 13 to 40, but had the largest equating biases at the low end of the score scale.

4.1.3 The Propensity Score Method: Predictions of Score Frequencies

One objective of this study was to assess how well the PSE methods, in comparison with other methods, predicted score frequencies of missing data. Better prediction of score frequencies of missing data is thought to reduce equating biases of the PSE method. This prediction assessment is useful in evaluating which collateral information set is more effective to be a stratification variable in the PSE method. Pearson chi-square statistics and likelihood ratio (LR) chi-square statistics were used to assess the closeness between the predicted frequencies and the true frequencies. Small Pearson chi-square statistics and likelihood ratio (LR) chi-square statistics indicate small differences between the predicted score frequencies predicted by collateral information

and the true (simulated) score frequencies. Note that, there are two missing parts when data are collected through the NEAT design—*Y* for population *P* and *X* for Population *Q*. These missing parts indicate that the sample from the population P was not administered to the test *Y* and that the sample from the population Q was not administered to the test *X*. This was not a conventional missing data typically occurring in survey research. These two statistics were therefore reported as measures of the prediction power for the missing data *Y* for the population P and the missing data *X* for the population *Q*, separately. The interpretation of chi-square and LR statistics is similar. For example, a chi-square statistics for *P* shows the degree to which score frequencies of missing data for examinees in the population *P* were predicted by a certain equating method, whereas a chi-square statistics for *Q* shows the degree to which score frequencies of missing data for examinees in the population *Q* were predicted by a certain equating method.

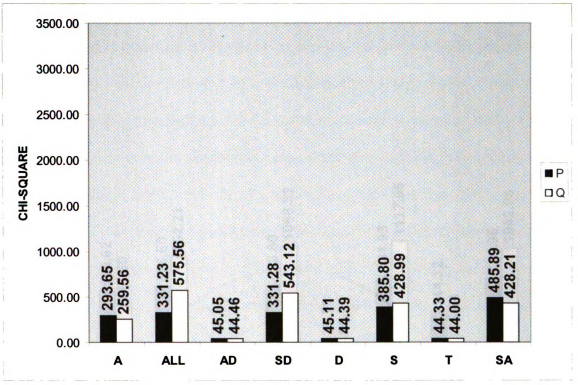

FIGURE 4.3a: PS chi-square statistics: long test and no group differences

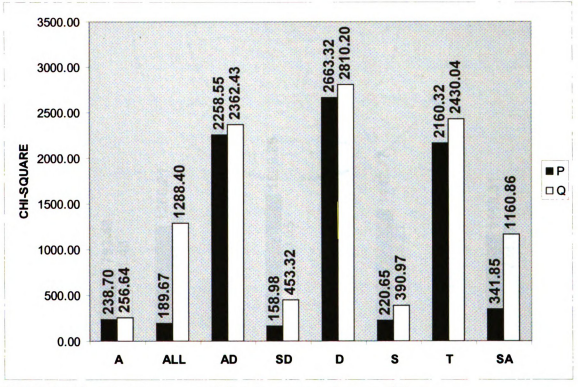

FIGURE 4.3b: PS chi-square statistics: long test and group differences

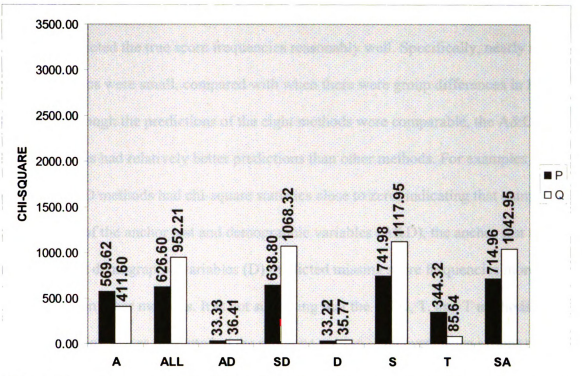

FIGURE 4.3c: PS chi-square statistics: short test and no group differences

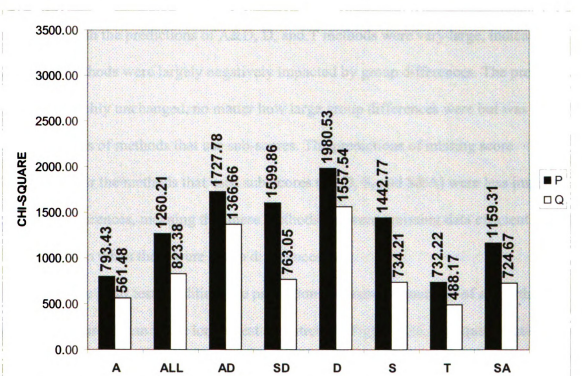

FIGURE 4.3d: PS chi-square statistics: long test and group differences

For the long test condition, when there were no group differences, all of the eight methods predicted the true score frequencies reasonably well. Specifically, nearly all chi-square statistics were small, compared with when there were group differences in Figure 4.2b. Even though the predictions of the eight methods were comparable, the A&D, T, and D methods had relatively better predictions than other methods. For examples, the A&D, T and D methods had chi-square statistics close to zero, indicating that using the combination of the anchor test and demographic variables (A&D), the anchor test true score (T), and demographic variables (D) predicted missing score frequencies more accurately than other methods. It is not surprising that the A&D, T, and T methods had smaller equating biases presented in the previous section. However, when groups differ greater in abilities, the predictions of some methods were worse than when there were no group differences, as seen in Figure 4.3b. Figure 4.3b shows that the chi-square statistics associated with the predictions of A&D, D, and T methods were very large, indicating that these methods were largely negatively impacted by group differences. The prediction of A was roughly unchanged, no matter how large group differences were but was below the predictions of methods that use sub-scores. The predictions of missing score frequencies for the methods that used sub-scores (S&D, S, and S&A) were less impacted by group differences, meaning that these methods recovered missing data efficiently in long tests even when there were group differences.

For the short test condition, the predictions of score frequencies of all methods were less accurate than in the longer test condition. In Figure 4.3c, chi-square statistics on average were about 500 for the sample P and 600 for the sample Q, which were larger than those chi-square statistics in Figure 4.3a. The common finding between Figure 4.3a

and Figure 4.3c is that the A&D, D, and T methods predicted score frequencies of missing data more accurately than other methods when there were no group differences. However, when groups differ greatly in abilities, all methods had less prediction power than when group differences were small. As see indicated in Figure 4.3d, the traditional PSE method (A) and the modified PSE method (T) better predicted score frequencies of missing data than other methods when there were group differences in the short test condition. The PSE method that used demographic variables had worse predictions of missing data of *P* and *Q* populations.

The second statistics used to assess the predictions of score frequencies of missing data was the likelihood ratio (LR) chi-square statistics. The interpretation of LR statistics is the same as the Pearson chi-square statistics, that is, smaller LR statistics indicate better predictions of score frequencies of missing data. Figures 4.4a to 4.4d present the LR statistics for different equating methods under different conditions. The results of LR statistics were consistent with the chi-square statistics. For example, when there was no group difference in the long test condition, the A&D, D, and T methods were the best in predicting score frequencies of missing data, as evidenced by smaller LR statistics in Figure 4.4a. But when group differences were greater, the methods that used sub-scores had better predictions, as shown by smaller LR statistics in Figure 4.4b.

Figure 4.4c shows that the A&D, D, and T methods were the best in predicting score frequencies of missing data in the short test condition when there were no group differences. This was evidenced by the fact that they had smaller LR statistics than other methods. However, when groups differ greatly in abilities, as seen in Figure 4.4d, none of the methods could predict score frequencies well. That is, they had large LR statistics.

Even though the methods that use sub-scores (S&D, S, and S&A methods) had larger LR statistics than other methods, they were not much impacted by group differences as compared to the methods that use demographic variables (A&D, and D). The predictions of the D and A&D methods were tremendously impacted by group differences as evidenced by lower LR statistics. The traditional PSE method (A) and the modified PSE method (T) were the two methods that best predicted score frequencies of missing data in the short test condition, when there were group differences.

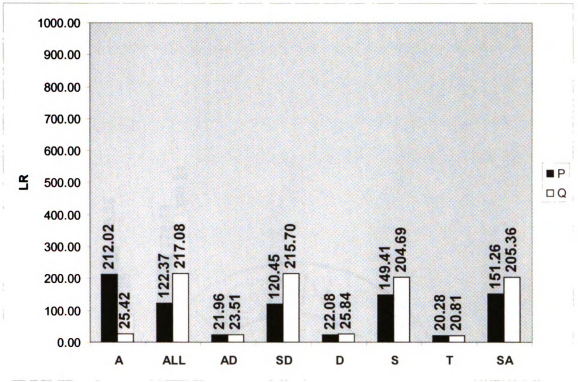

FIGURE 4.4a: PS likelihood ratio (LR) statistics: long test and no group differences

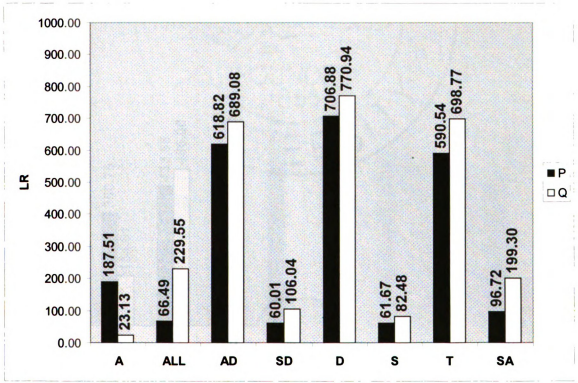

FIGURE 4.4b: PS likelihood ratio (LR) statistics: long test and group differences

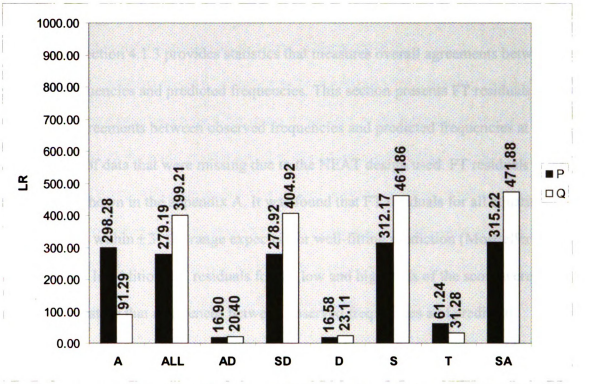

FIGURE 4.4c: PS likelihood ratio (LR) statistics: short test and no group differences

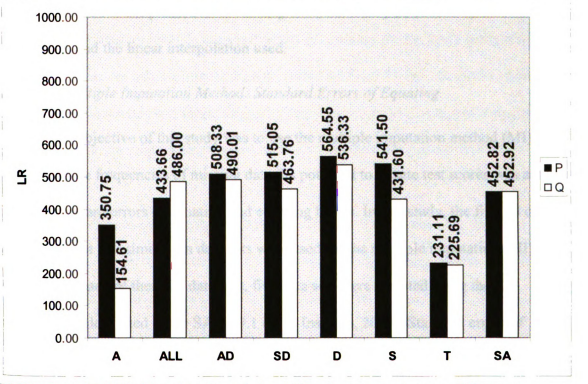

FIGURE 4.4d: PS likelihood ratio statistics (LR): short test and group differences

4.1.4 The Propensity Score Method: FT Residuals

The Section 4.1.3 provides statistics that measures overall agreements between observed frequencies and predicted frequencies. This section presents FT residuals illustrating agreements between observed frequencies and predicted frequencies at the score point x of data that were missing due to the NEAT design used. FT residuals were graphed and shown in the appendix A. It was found that FT residuals for all equating methods were within ± 3, the range expected for well-fitting prediction (Mostteller & Youtz, 1961). In addition, FT residuals for the low and high ends of the score were close to zero, suggesting that differences between observed frequencies and predicted frequencies are very small at the high and low ends of the scale. This gives a clue that poor estimates of standard errors of equating and equating biases at the bottom end of the scale are not due to the prediction of missing data, but it may be caused by zero frequencies and the linear interpolation used.

4.1.5 The Multiple Imputation Method: Standard Errors of Equating

One objective of this study was to use the multiple imputation method (MI) to compute score frequencies of missing data. Its potential to equate test scores was assessed through standard errors of equating and equating biases. In this study, the first 20 data sets of the first 100 simulation data sets were used for the multiple imputation (MI) method. For each of these 20 data sets, five data sets were imputed using the EM algorithm implemented in the SAS V9.1 (SAS Institute, 2003). Standard errors of equating and equating biases were computed based on the five imputations and the resulting standard errors were averaged across 20 data sets. The averages were used as

the value reflecting "standard errors of equating" of the multiple imputation method. Equating biases were computed and averaged in the same way.

The MI standard errors of equating were graphically presented in Figures 4.5a – 4.5d. When there were no group differences in abilities in the long test condition, Figure 4.5a shows that, at the score scale from 22 to 48, all equating methods produced very small standard errors of equating, that is, standard errors of equating produced by all of the eight methods were close to zero. However, at the low and high ends of the score scale, standard errors of equating of all methods were more different, but the differences were greater at the low end of the score scale than at the high end of the score scale. When groups of examinees differed greatly in abilities, Figure 4.5b shows that all equating methods had similar small standard errors of equating at score points between 23 and 55, except for the PSE method that uses the combination of sub-scores and demographic variables (S&D). Standard errors of equating for the S&D method were larger than those for other methods at score points between 23 and 60. Standard errors of equating of all methods had larger degree of fluctuations at the low end of the score scale, suggesting that equating functions were unreliably produced at the low end of the score scale.

When there were no group differences in terms of abilities in the short test condition and, Figure 4.5c shows that all methods produced similar standard errors of equating but larger differences were found the low end of the score scale. Specifically, at the score points between 14 and 40, standard errors produced by all methods were close to zero, meaning that all of these methods produced equating functions that had small degree of equating errors at these score points. But all equating methods had more

divergent equating errors at the low end of the score scale. Even though all methods produced similar standard errors of equating, the traditional PSE method (T) and the PSE method that uses the combination of sub-scores and anchor test (S&A) had the smallest standard errors of equating.

It was found in Figures 4.5c and 4.5d that MI methods had smaller standard errors of equating, implying that the variation of equated scores among the five imputations was small. Moreover, MI methods had small standard errors of equating, regardless of the group differences.

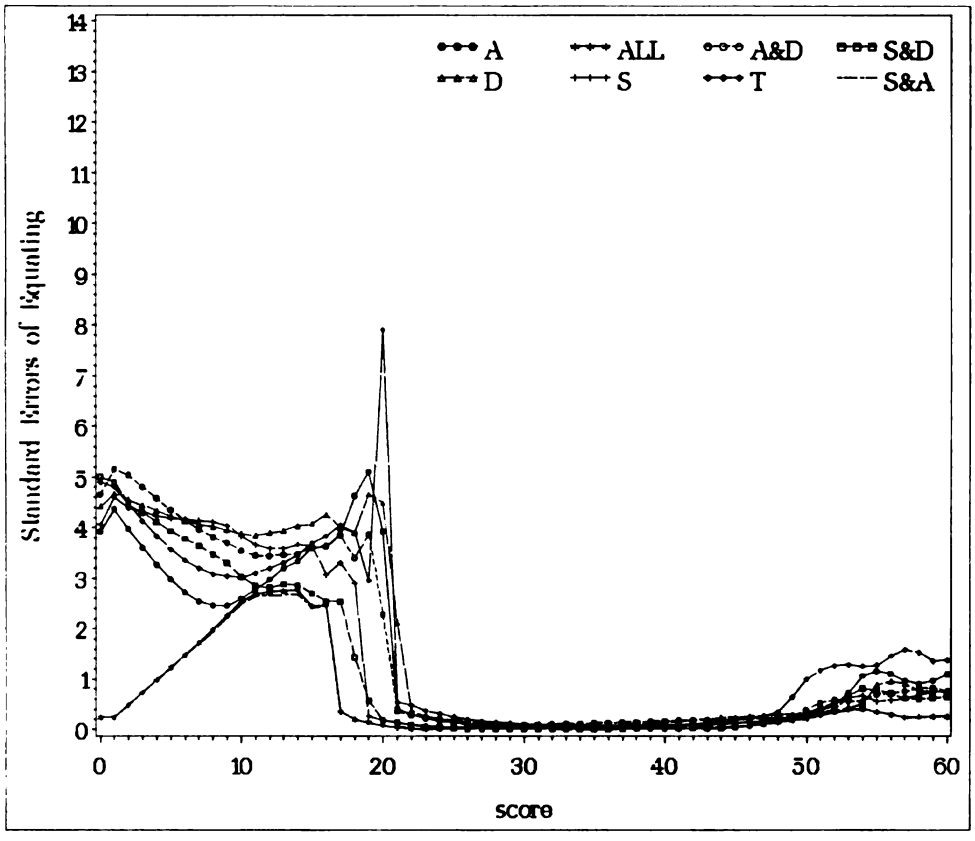

FIGURE 4.5a: MI standard errors of equating: long test and no group differences

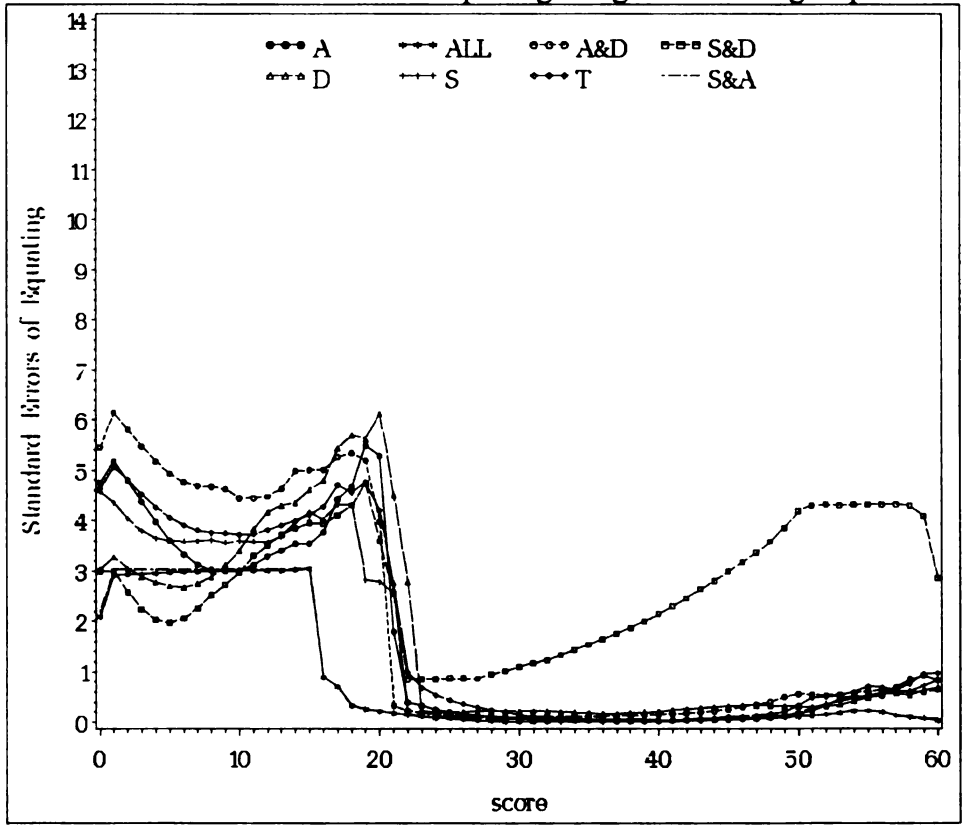

FIGURE 4.5b: MI standard errors of equating: long test and group differences

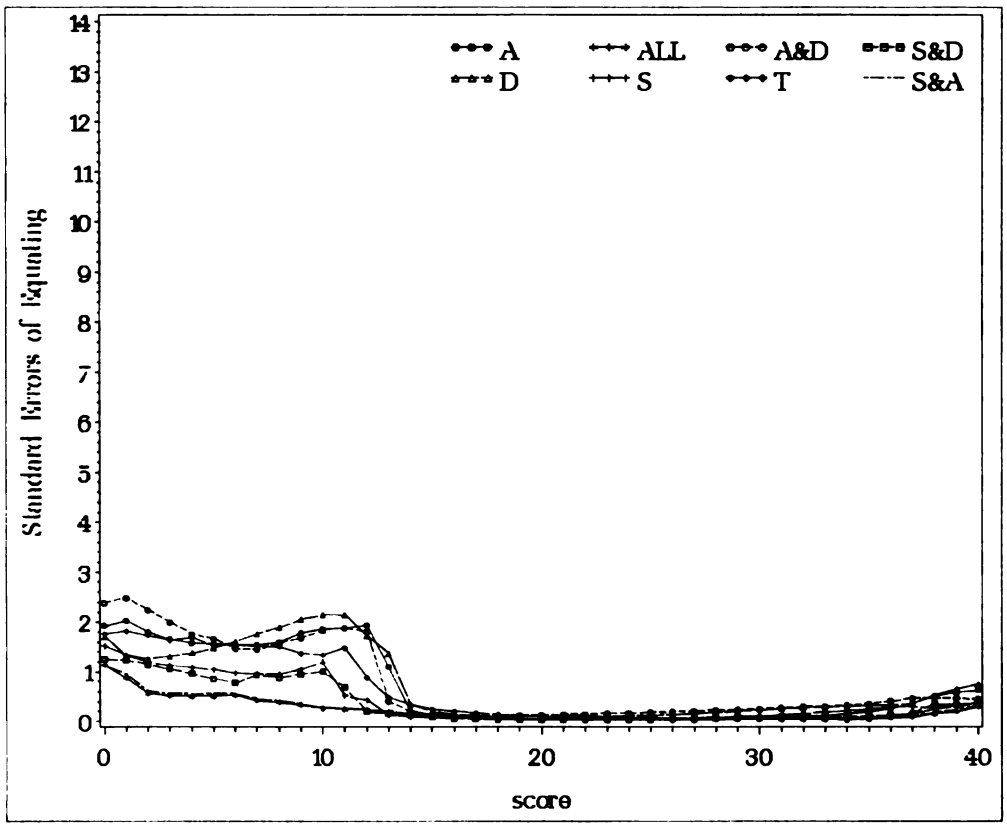

FIGURE 4.5c: MI standard errors of equating: short test and no group differences

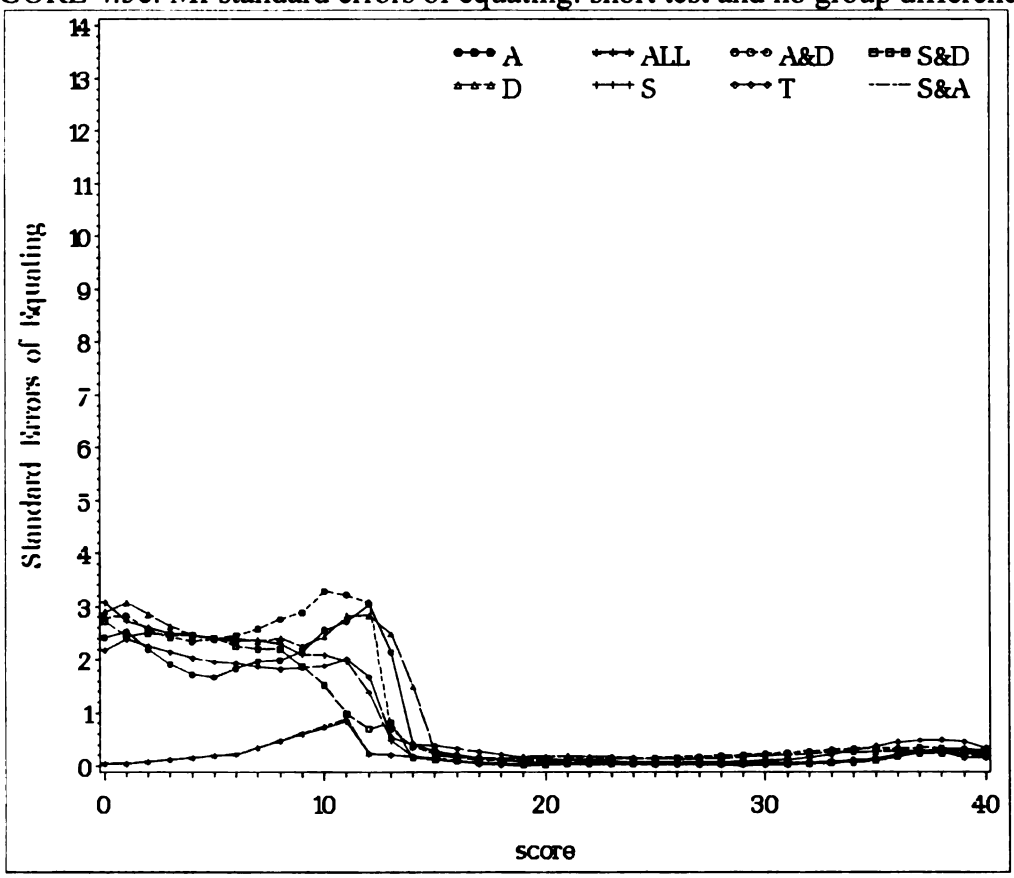

FIGURE 4.5d: MI standard errors of equating: long test and group differences

In the condition where tests were shorter and groups differed greatly in abilities, it was found that the standard errors of equating of all MI equating methods were not much affected by the group differences. The standard errors of equating in Figure 4.5d were comparable to those standard errors of equating in Figure4.5c. That is, all equating methods had smaller standard errors of equating at the score points from 15 to 40, but standard errors of equating were larger at the low end of the score scale. As noted earlier, the PSE method that uses the anchor test true score (T) and the PSE method that uses sub-scores and the anchor test score (S&A) had the smallest standard errors of equating.

4.1.6 The Multiple Imputation Method: Equating Bias

Equating biases indicate how large the equating function deviated from the criterion equating function. A criterion equating function was obtained through the equating of the complete (simulated) data. Specifically, the criterion equating function was obtained by equating scores of the simulation data which are data without missing data or the pseudo-test data. The resulting equating function was called "the criterion equating function." Then test data with missing data were generated as if these data were collected through the NEAT design. A series of test score equating was performed using the generated data. The resulting equating function was then compared to the criterion equating function. The resulting difference was defined as the equating bias. Figures 4.6a – 4.6d compare equating biases produced by different equating methods under different conditions. As noted earlier, when the MI method was investigated, equating biases presented in this section were the averages of the 20 data sets.

Figure 4.6a shows equating biases of the condition in which test length was 60 and there were group differences. The PSE method that uses demographic variables, the

modified PSE method (T), and the traditional PSE method (A) underestimated equating functions as indicated by negative equating biases. The PSE method that uses the combination of anchor test and demographic variables (A&D) overestimated equating functions as seen by positive equating biases. The PSE method that uses the combination of sub-scores and demographic variables (S&D), the PSE method that uses all collateral information (ALL), the PSE method that uses sub-scores (S), and the PSE method that uses the combination of sub-scores and anchor test (S&A) overestimated and underestimated equating functions, depending on the score points on the score scale. In general, the S&D, S, ALL, and S&A methods had smaller equating biases than other methods. Note that these methods involve the uses of subs-cores.

Figure 4.6b shows that when the test length was 60 and groups differed greatly in abilities, all equating methods tended to overestimated equating functions, as indicated by the fact that the equating biases shifted upward as compared to the Figure 4.6a. But this pattern was not true for the S&D method at the low end of the score scale. This pattern shows that the equating biases for the D and T methods that had negative biases in Figure 4.6a were shifted upwards to somewhere close to zero, when groups differed in abilities. However, the methods that had positive equating biases in Figure 4.6a had even larger equating biases when groups differ greater in abilities. When groups differ greatly in ability, the D and T method had smaller equating biases than other methods.

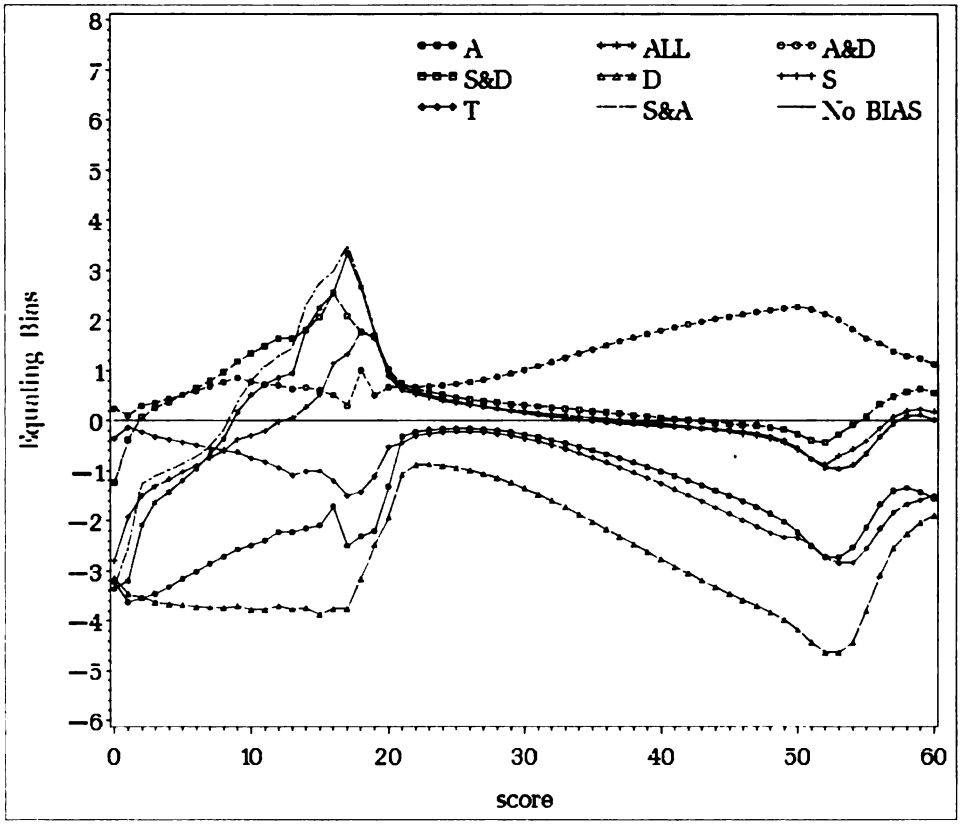

FIGURE 4.6a: MI equating biases: long test and no group differences

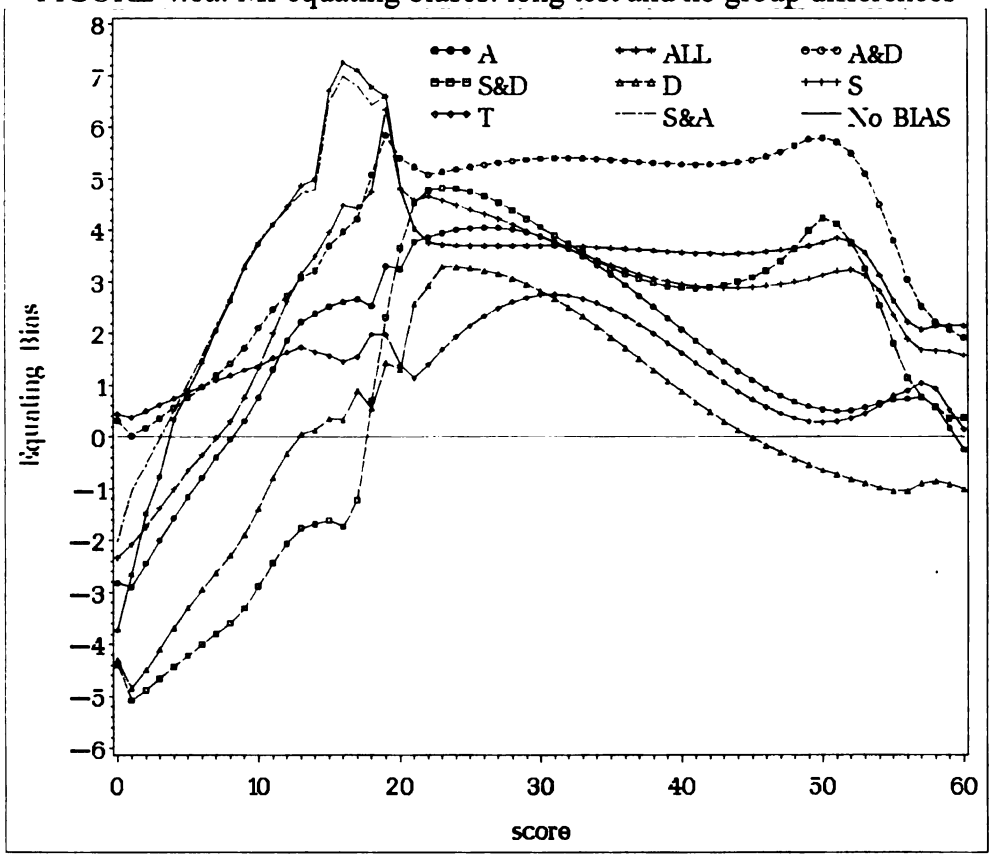

FIGURE 4.6b: MI equating biases: long test and group differences

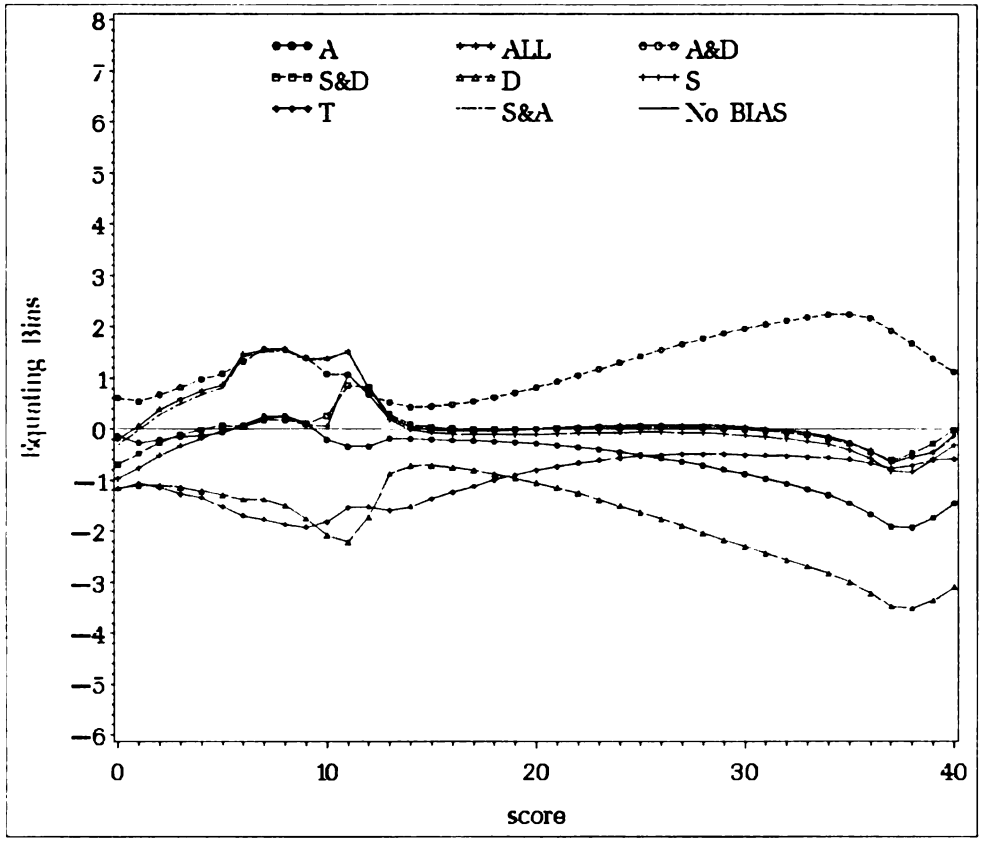

FIGURE 4.6c: MI equating biases: short test and no group differences

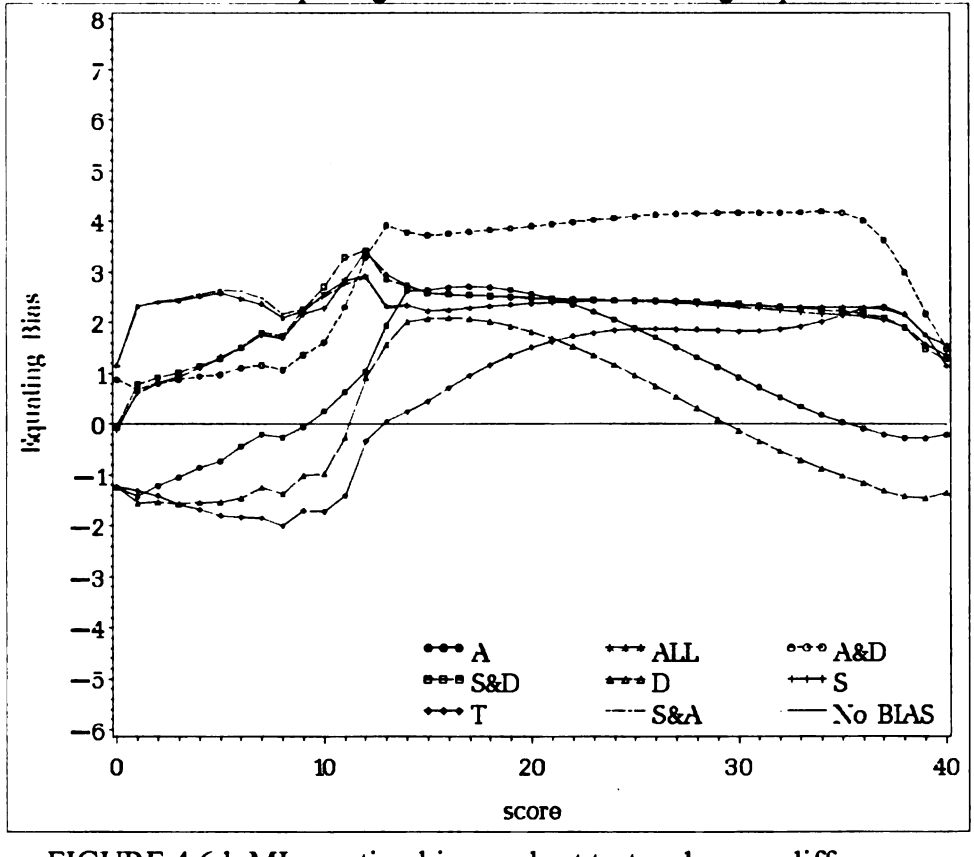

FIGURE 4.6d: MI equating biases: short test and group differences

Figure 4.6c presents equating biases when test length was 40 and there were no group differences. As seen in this Figure, the PSE method that uses demographic variables (D) and the method that uses anchor test true score (T) underestimated equating functions, as evidenced by negative equating biases. The method that uses the combination of anchor test score and demographic variables (A&D) overestimated equating functions as indicated by positive equating biases. The PSE method that uses sub-score (S) and the PSE method that uses the combination of sub-sores and demographic variables (S&D) had the smallest equating biases which are close to zero. The PSE method that uses all collateral information (ALL) and the PSE method that uses the combination of sub-scores and anchor test score (S&A) had comparable equating biases. It is clear that the ALL, S, S&A, and S&D methods had comparable equating biases at the score points ranging from 14 to 40. The traditional PSE method (A) had small negative biases but larger than those produced by ALL, S, S&A, and S&D. But at the low end of the score scale, the traditional PSE method (A) had the smallest equating biases.

Figure4.6d shows equating biases when test length was 40 and there were group differences. Similar to the Figure4.6b, when groups differ in abilities all equating methods tended to overestimate equating functions, as evidenced by larger degree of positive equating errors as compared to those in the Figure 4.6c. In this condition, the modified PSE method (T) and the PSE method that uses demographic variables (D) were two methods that had the smallest equating biases.

4.1.7 The Multiple Imputation Method: Prediction of Score Frequencies of Missing Data

One objective of this study is to investigate the efficiency of the multiple imputation method in predicting score frequencies of missing data. As noted previously, better score frequencies of missing data results in smaller equating biases for the PSE method. Therefore, the method that best predicts score frequencies of missing data would have the smallest equating biases. Like the propensity score method section, two statistics used in this study to measure the closeness between the predicted frequencies and the true (simulated frequencies) included Pearson chi-square statistics, and likelihood ratio (LR) chi-square statistics. The smaller chi-square, and LR statistics indicate the greater degree of closeness between the two frequencies. Figures 4.7a - 4.7d show chi-square statistics and Figures 4.8a – 4.8d present LR statistics. Note that in these Figures, goodness-of fit statistics were presented separately for different two populations (*P* and *Q*). For example, a chi-square statistics for *P* shows the degree to which score frequencies of missing data for examinees in the population *P* were predicted by a certain equating method, whereas a chi-square statistics for *Q* shows the degree to which score frequencies of missing data for examinees in the population *Q* were predicted by a certain equating method.

Figure 4.7a compares chi-square statistics for different equating methods when the test length was 60 and there were no group differences. It was obvious that the PSE method that uses all collateral information (ALL), the PSE method that uses sub-scores (S), the method that uses the combination of sub-scores and demographic variables (S&D), the modified PSE method (T), and the method that uses the combination of the anchor test score and sub-scores were combined (S&A) had more comparable and smaller chi-square statistics than other methods. The traditional PSE method (A) had total

chi-square statistics larger than the modified PSE method (T). The PSE method that uses demographic variables (D) had the largest chi-square.

Figure4.7b compares chi-square statistics when the test length was 60 and there were group differences. It was obvious that when groups differ in abilities, nearly all equating methods had larger chi-square statistics, except for the method that uses demographic variable (D), and the increments in chi-square statistics were more evidenced for the sample of population *P* or the population that had lower abilities. It was also obvious that all equating methods predicted score frequencies of missing data better for *Q* than *P*, as indicated by smaller chi-square statistics for *Q*. In other words, all equating methods better predicted frequencies for the population that had higher abilities than the population that had lower abilities. The method that uses the combination of the anchor test and demographic variables (A&D) had the largest total chi-square, while the modified PSE method (T) and the PSE method that uses demographic variables (D) had smaller total chi-squares than other methods. This result was consistent with the equating bias results in that the T and D methods were two methods that had smaller equating biases.

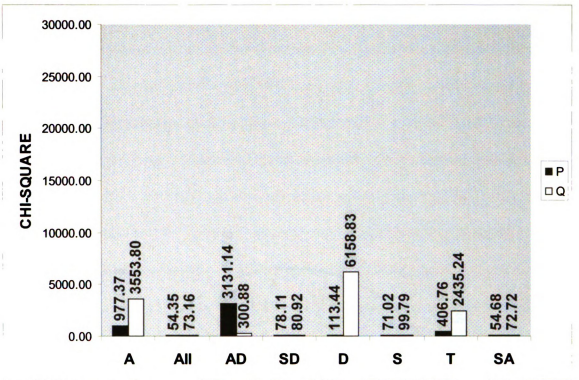

FIGURE 4.7a: MI chi-square statistics: long test and no group differences

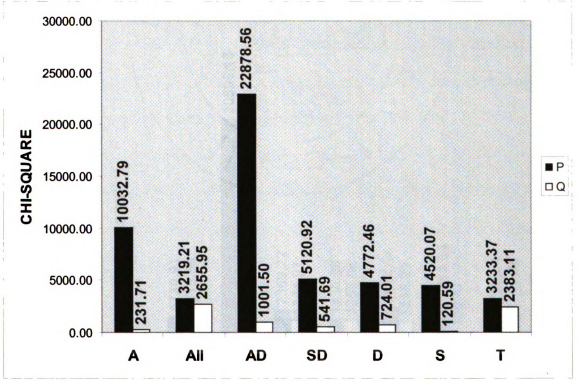

FIGURE 4.7b: MI chi-square statistics: long test and Group Differences

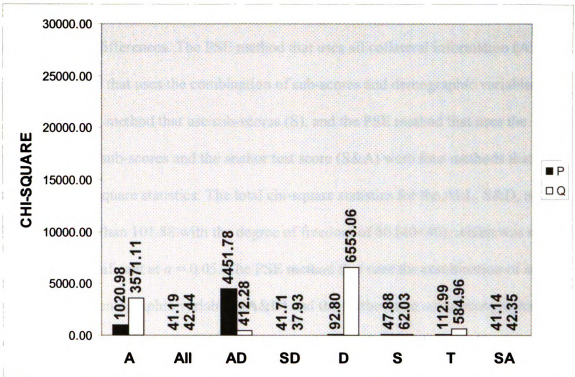

FIGURE 4.7c: MI chi-square statistics: short test and no group differences

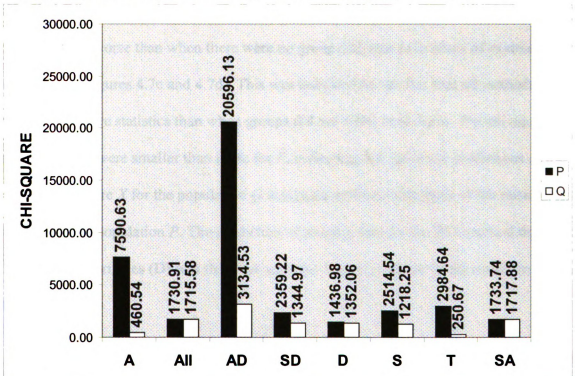

FIGURE 4.7d: MI chi-square statistics: short test and group differences

Figure4.7c compares chi-square statistics when the test length was 40 and there were no group differences. The PSE method that uses all collateral information (ALL), the PSE method that uses the combination of sub-scores and demographic variables (S&D), the PSE method that use sub-scores (S), and the PSE method that uses the combination of sub-scores and the anchor test score (S&A) were four methods that had small total chi-square statistics. The total chi-square statistics for the ALL, S&D, and S&A were less than 101.88 with the degree of freedom of 80 (40+40), which was not statistically significant at $\alpha = 0.05$. The PSE method that uses the combination of anchor test score and demographic variables (A&D) and the method that uses demographic variables (D) had the largest chi-square statistics.

Figure 4.7d compared chi-square statistics when the test length was 40 and groups differed in abilities. When groups differ greatly in abilities, the predictions of all equating methods were worse than when there were no group differences in terms of examinees' abilities (see Figures 4.7c and 4.7d). This was indicated by the fact that all methods had higher chi-square statistics than when groups did not differ in abilities. The chi-square statistics for Q were smaller than those for P, indicating that again the predictions of the missing test score X for the population Q was more accurate than those of the missing test score Y for the population P. The prediction of missing data for the PSE method that uses demographic variables (D) was the most accurate, which is similar to the results in Figure 4.7b.

The second statistics used to measure the closeness of the predicted frequencies and the true (simulated) frequencies was the likelihood ratio (LR) chi-square statistics. Figures 4.8a-4.8d presents LR statistics for different equating methods under different

simulation conditions. The interpretation of these statistics is the same as that of chi-square statistics. The results shows similar pattern of predictions between chi-square statistics and LR statistics. For example, predictions of score frequencies were found to be more accurate when there were no group differences, and the PSE method that uses all collateral information (ALL), the PSE method that uses the combination of sub-scores and demographic variables (S&D), the PSE method that uses sub-scores (S), and the PSE method that uses the combination of sub-scores and the anchor test score (S&A) had better predictions of score frequencies of missing data, as evidenced by smaller total chi-squares for both *P* and *Q*. Figure 4.8b and Figure 4.8d show that the predictions for the population *Q* were better than that for the population *P*, and all equating methods had negative impacts from group differences, except for the method that uses demographic variables.

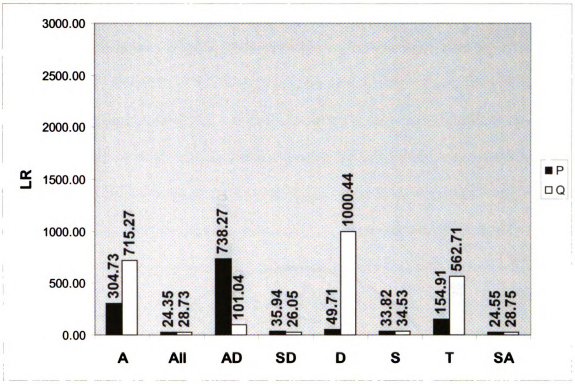

FIGURE 4.8a: MI likelihood ratio (LR) chi-square statistics: long test and no group differences

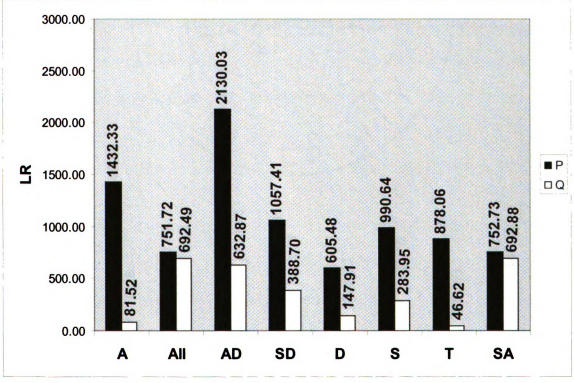

FIGURE 4.8b: MI likelihood ratio (LR) chi-square statistics: long test and group differences

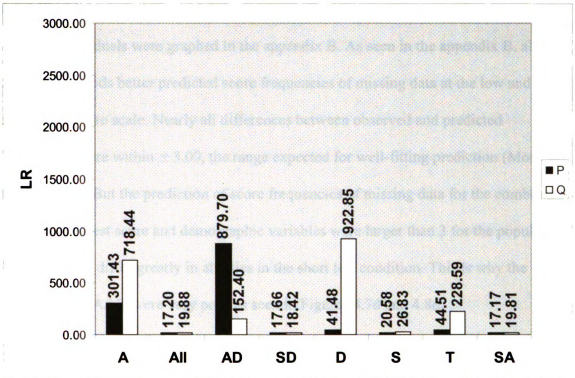

FIGURE 4.8c: MI likelihood ratio (LR) chi-square statistics: short test and no group differences

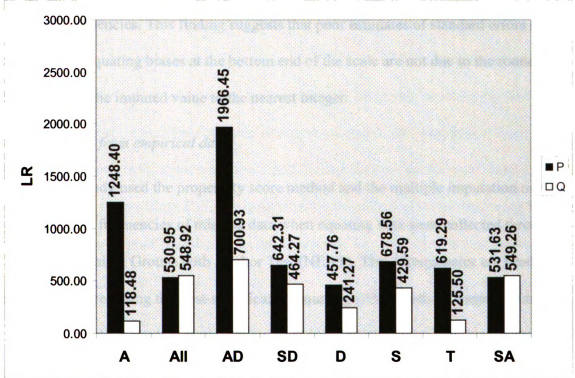

FIGURE 4.8d: MI likelihood ration (LR) chi-square statistics: short test and group differences

4.1.8 The Multiple Imputation Method: FT Residuals

FT residuals were graphed in the appendix B. As seen in the appendix B, all equating methods better predicted score frequencies of missing data at the low and high ends of the score scale. Nearly all differences between observed and predicted frequencies were within ± 3.00, the range expected for well-fitting prediction (Mosteller, Youtz, 1961). But the prediction of score frequencies of missing data for the combination of the anchor test score and demographic variables were larger than 3 for the population P, when group differ greatly in abilities in the short test condition. This is why the predictions of A&D were very poor as seen in Figures 4.7d and 4.8d.

The FT residuals in the appendix B also show that the rounding effect is trivial for the multiple imputation method at the low and high ends of the score scale. That is, it is found that nearly all equating methods had very small differences between observed and predicted frequencies. This finding suggests that poor estimates of standard errors of equating and equating biases at the bottom end of the scale are not due to the rounding used to round the imputed value to the nearest integer.

4.2 The results from empirical data

This study used the propensity score method and the multiple imputation method to obtain score frequencies of missing data when equating data were collected through the Non-Equivalent Groups with Anchor Test (NEAT). These frequencies are needed to equate test scores using the post-stratification equating (PSE) method. Empirical data used in this study to explore these two methods is the Mathematics Teaching in the 21st Century data which is a small scale international study aiming at measuring mathematics and mathematic pedagogy knowledge of teacher candidates or future math teachers of 6

countries. There are two test forms *X* and *Y*, having 38 and 30 items, respectively, plus 24 common items. *X*-scores were equated to *Y*-score using the same methods as for the simulation study. For the propensity score method, this study used the Kernel Equating (KE) software (Chen, Yan, Hemat, Han, & von Davier, 2008) because the KE software can estimate standard errors of equating. However, for the multiple imputation method, the SAS program developed and used in the simulation section was used to estimate standard errors of equating, which were the averages of standard errors of equating across 5 imputations. The equating biases were computed as the differences between the resulting equated scores and the criterion equated scores, where the criterion was obtained from the Item Response Theory (IRT) true score equating.

There were two conditions manipulated to vary group differences in terms of examinees' abilities on the quality of equating. The first condition explored standard errors of equating and equating biases when there were no group differences. This condition was achieved by using all cases in the MT21 data. The second condition compared standard errors of equating and equating biases among equating methods when groups differ in abilities. This condition was manipulated by deleting data of two high performing countries from the test form 1 and deleing data of two low performing countries from the test form 2. The manipulation resulted in that the group taking the test form 2 performed better than the group taking the test form 1. The next section presents standard errors of equating and equating biases of different PSE equating methods.

4.2.1 The Propensity Score Method: Standard Errors of Equating

Figures 4.9a-4.9b compare standard errors of equating when the propensity score method was used. Specifically, Figure 4.9a presents standard errors of equating when

groups were identical in terms of examinees' abilities, while Figure 4.9b presents standard errors of equating when groups differ in abilities. It was obvious that when groups were identical, the traditional PSE method (A) and the modified PSE method (T) had smaller standard errors of equating. The PSE method that uses sub-scores (S), the PSE method that uses the combination of sub-scores and demographic variables (S&D), and the PSE that uses the combination of all collateral information (ALL) had the largest standard errors of equating. However, although all equating methods had different standard errors of equating, their standard errors of equating were small.

When groups differed in abilities, all methods had even more different standard errors of equating but the modified PSE method (T) had the smallest standard errors of equating which were less than 0.5, followed by the traditional PSE method (A). Other equating methods had comparable standard errors of equating, which were larger at the low and high ends of the score scale but smaller at the middle of the score scale.

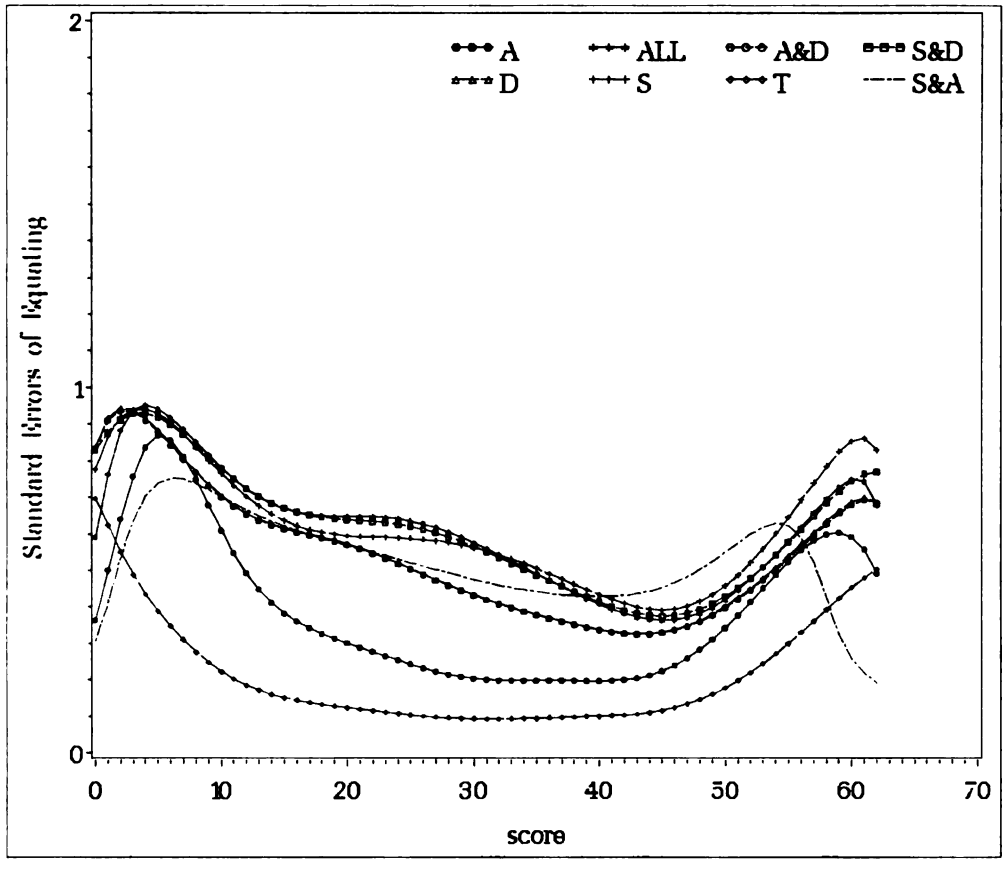

FIGURE 4.9a: PS standard errors of equating: no group differences

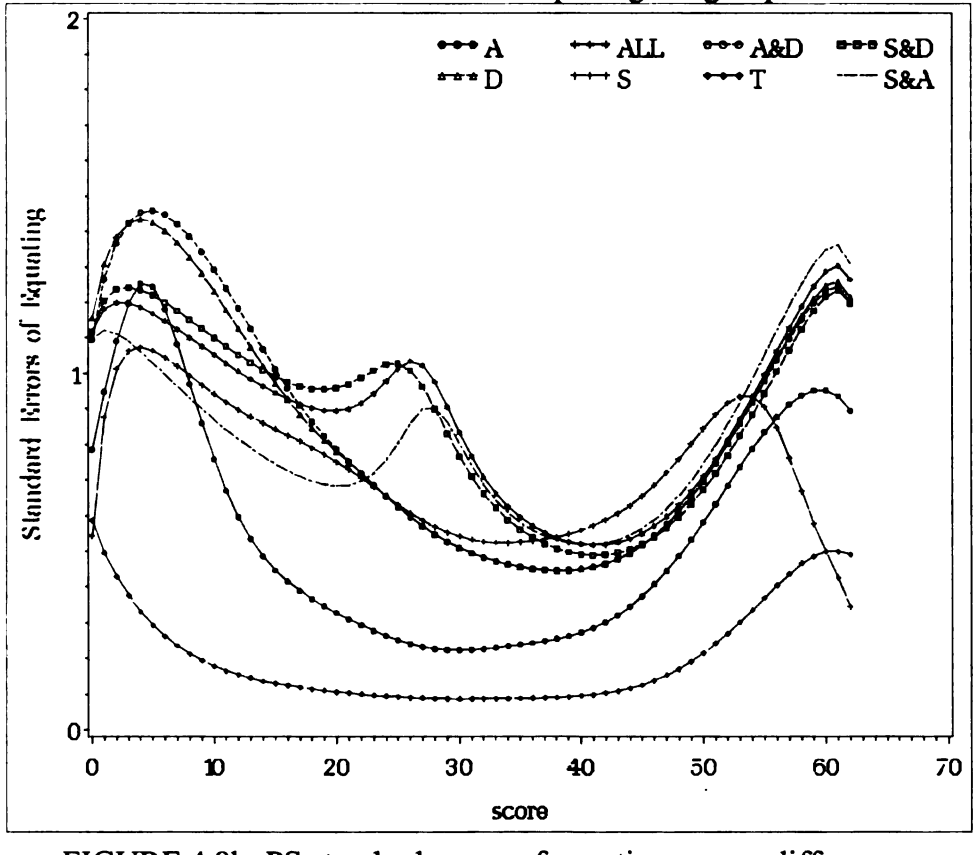

FIGURE 4.9b: PS standard errors of equating: group differences

4.2.2 The propensity Score Method: Equating Biases

Figure 4.10a displays equating biases when there were no ability differences between groups of examinees, while Figure 4.10b shows equating biases when there were ability differences between groups of examinees. As seen in Figure 4.10a, the PSE method that uses demographic variables (D) and the PSE method that uses the combination of the anchor test and demographic variables (A&D) had smaller equating biases than other methods, followed by the traditional PSE method (A) and the modified PSE method (T). The modified PSE method (T) had large positive equating biases at the low end of the score scale and negative equating biases at the high end of the score scale. The traditional PSE method (A) had large negative equating biases at the low end of the score scale. The PSE method that uses the combination of all collateral information (ALL), the PSE method that uses the combination of sub-scores and demographic variables(S&D), the PSE method that uses the combination of sub-scores and anchor test (S&A), and the PSE method that uses the sub-score (S) all had large negative equating biases. This might be due to the issue of test length discussed in the next chapter.

When groups differ in ability, as seen in Figure 4.10b, nearly all equating methods had large negative biases. The PSE method that uses sub-scores (S) had the smallest equating biases. Equating biases shifted downward to be more negative values, as compared to Figure 4.10a. However, it was not the case for the modified PSE method (T). The modified PSE method (T) method had even larger equating biases when there were group differences. Specifically, at the low end of the score scale, the equating biases were large and positive with a maximum of 13.00, while large negative equating biases were found at the high end of the score scale with a maximum of -12.00.

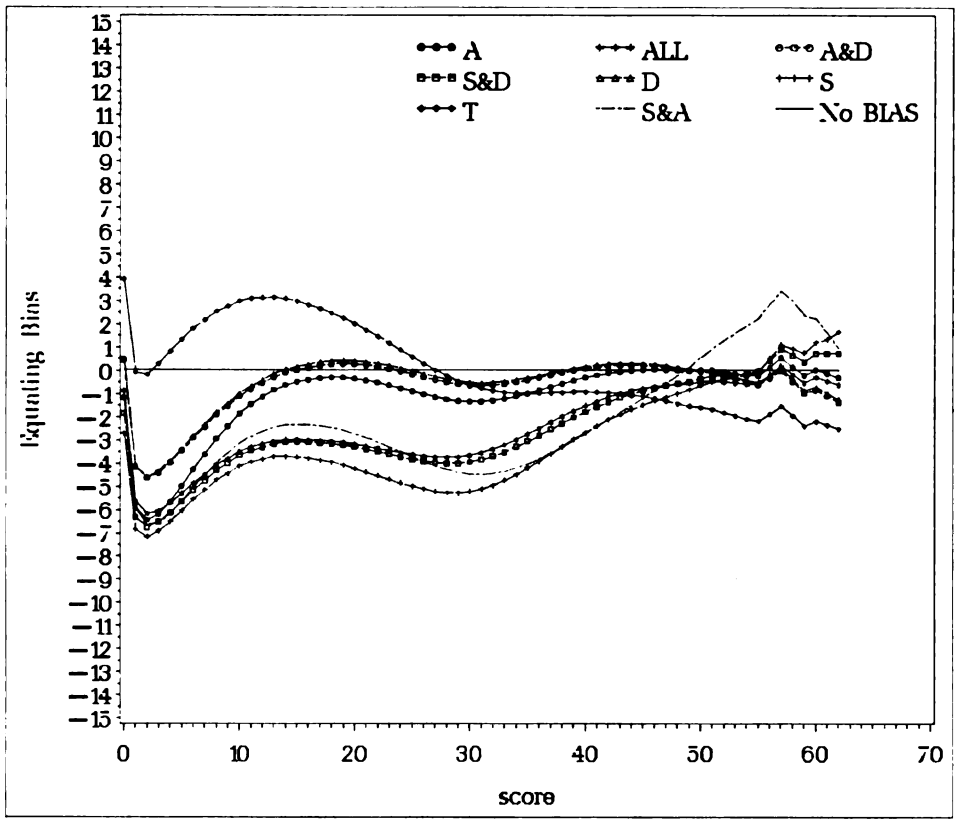

FIGURE 4.10a: PS equating biases : no group differences

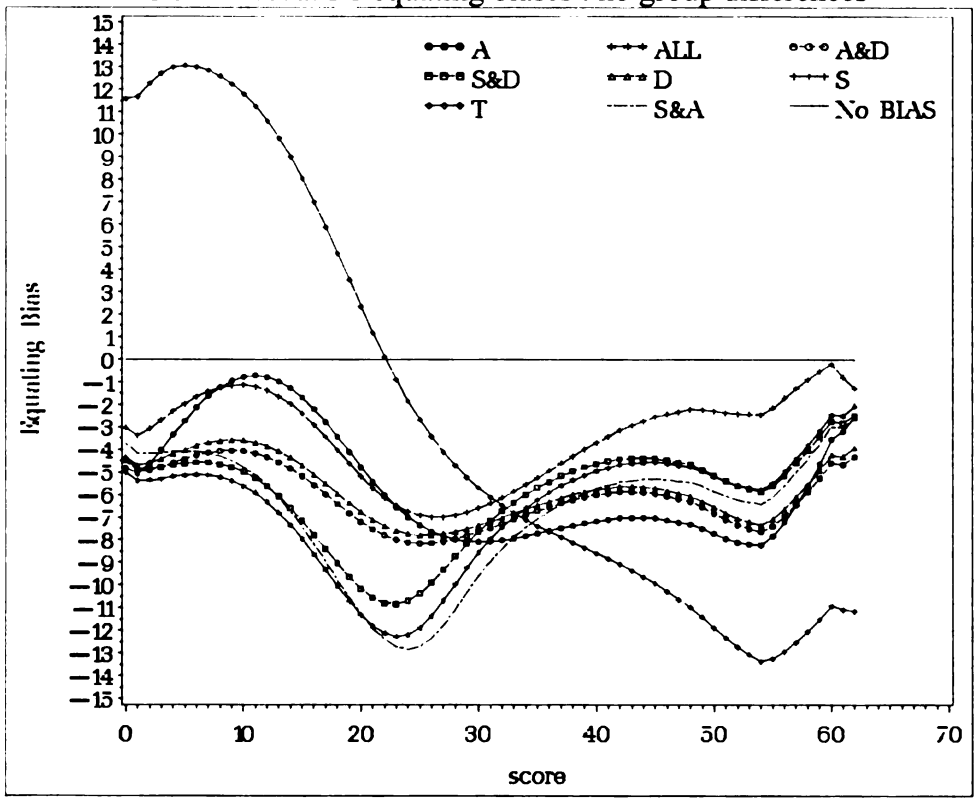

FIGURE 4.10b: PS equating biases : Group Differences

The equating biases of the S method tended to be unchanged from Figure 4.10a to Figure 4.10b, indicating that the S method had biased equating functions but was more tolerance to ability differences than other methods. Since all equating methods shifted their equating biases downwards but the S method did not get much impact from group differences, the equating biases for the S method were more close to zero than other methods. Although the S method had relatively smaller equating biases, their equating biases at the low end of the score scale were comparable to those equating biases of the traditional PSE method at the same range of score points. It can be noted from the Figure 4.10b that the PSE methods that use sub-scores as the information to impute missing data all had large negative biases.

4.2.3 The Multiple Imputation Method: Standard Error of Equating

The multiple imputation (MI) method was another method used in this study to explore its feasibility to compute the frequencies of missing data of the NEAT design. Figure 4.11a and Figure 4.11b compare its standard errors of equating under two conditions. Specifically, Figure 4.11a displays the standard errors of equating when there were no group differences in terms of examinees' abilities, while the Figure 4.11b presents standard errors of equating when groups differed in abilities.

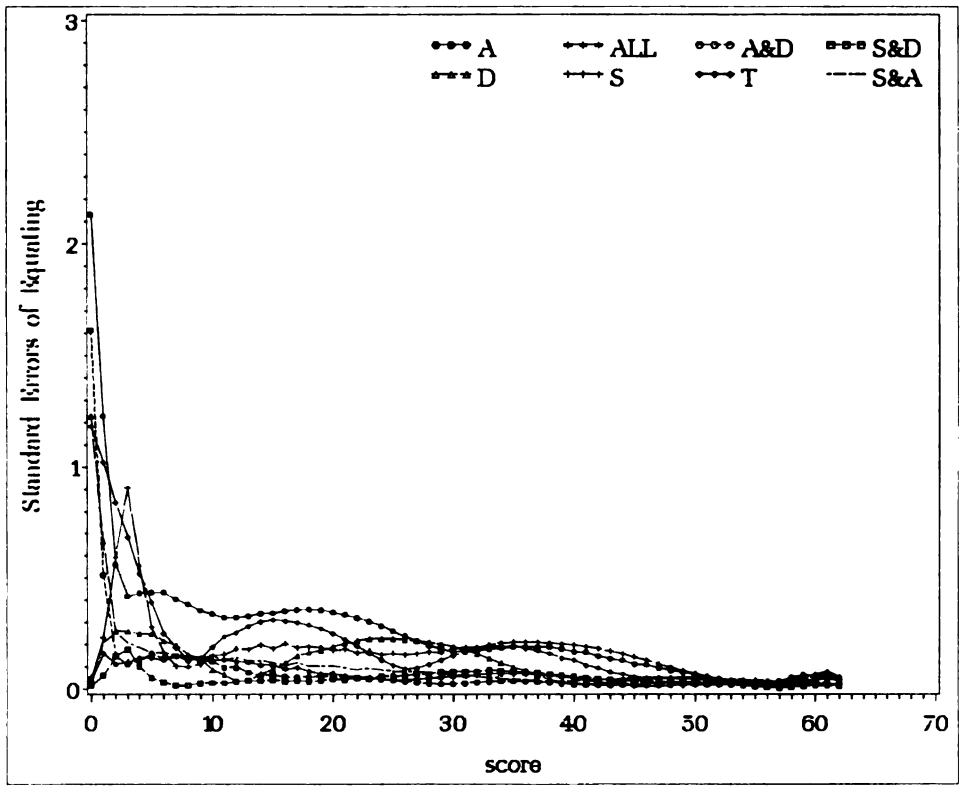

FIGURE 4.11a: MI standard errors of equating: no group differences

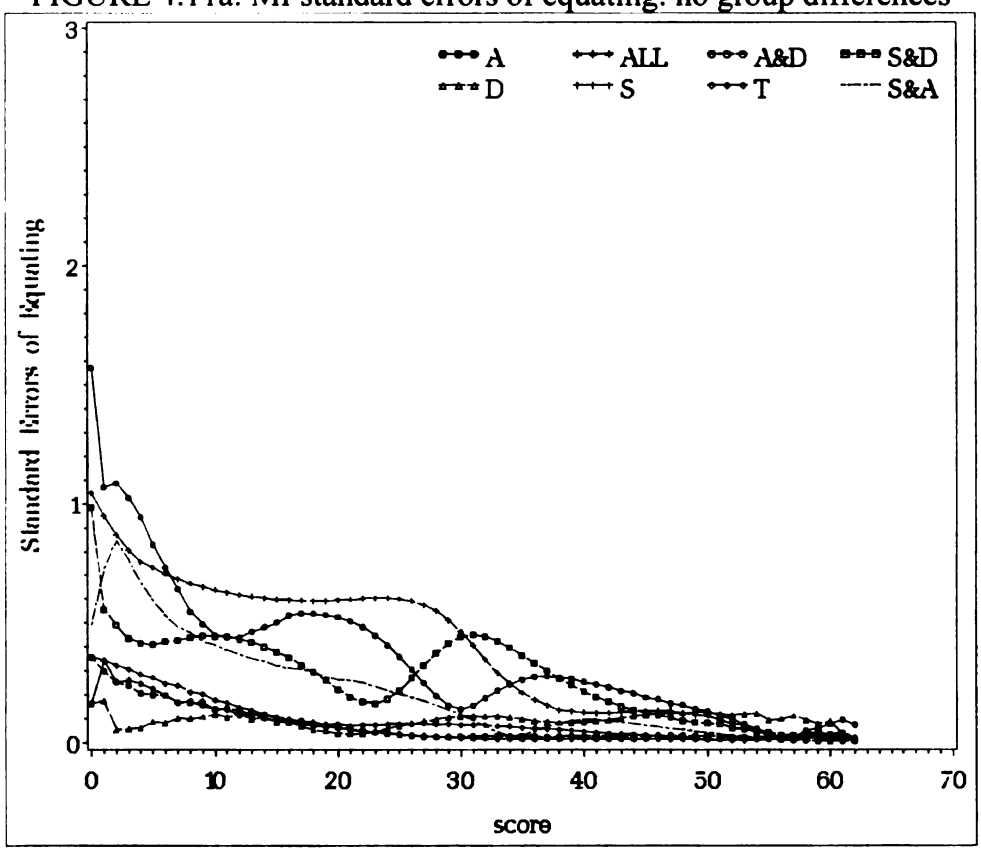

FIGURE 4.11b: MI standard errors of equating: group differences

As seen in Figure 4.11a, all equating methods had small standard errors of equating approximately less than .5 at the score points from 5 to 62. All equating methods had large standard errors of equating at the low end of the score scale, especially for the traditional PSE method (A). Even though all methods had comparable standard errors of equating, the PSE method that uses the combination of sub-scores and demographic variables (S&D) and the method that uses the demographic variables (D) had approximately smaller standard errors of equating than other methods.

In Figure 4.11b, when groups differ greatly in abilities, all equating methods had more different standard errors of equating. Standard errors of equating of the S, A, S&D, and S&A methods were larger than when there were no ability differences between groups of examinees. However, the standard errors of the D, A&D, and T, and ALL methods were smaller than other methods and tended to be unchanged from Figure 4.11a to Figure 4.11b.

4.2.4 The Multiple Imputation Method: Equating biases

Figures 4.12a and 4.12b compare equating biases of different PSE equating methods that used the multiple imputation (MI) method to compute score frequencies of missing data. Specifically, Figure 4.12a displays equating biases when there were no group differences in terms of examinees' abilities, while Figure 4.12b displays equating biases of different equating methods when there were group differences. Note that the criterion equating function used to compute equating biases in this study was the equating function obtained through the IRT true score equating.

When there were no group differences, equating methods that had smaller equating biases close to zero included the PSE method that uses demographic variables

(D), the PSE method that uses the combination of the anchor test and demographic variables (A&D), and the traditional PSE method (A). The modified PSE method (T) had large positive biases, while the PSE method that uses sub-scores (S) and the method that uses the combination of the sub-scores and demographic variables (A&D) had both positive and negative biases depending on the location on the score scale. The method that uses all collateral information tended to have negative equating biases.

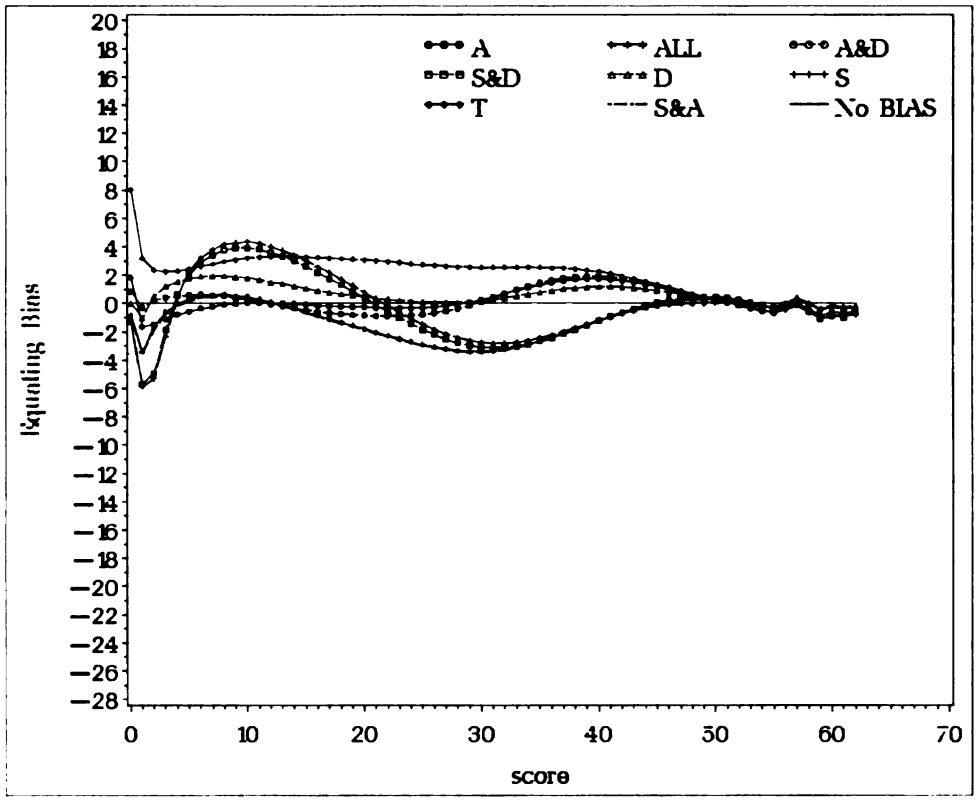

FIGURE 4.12a: MI equating biases: no group differences

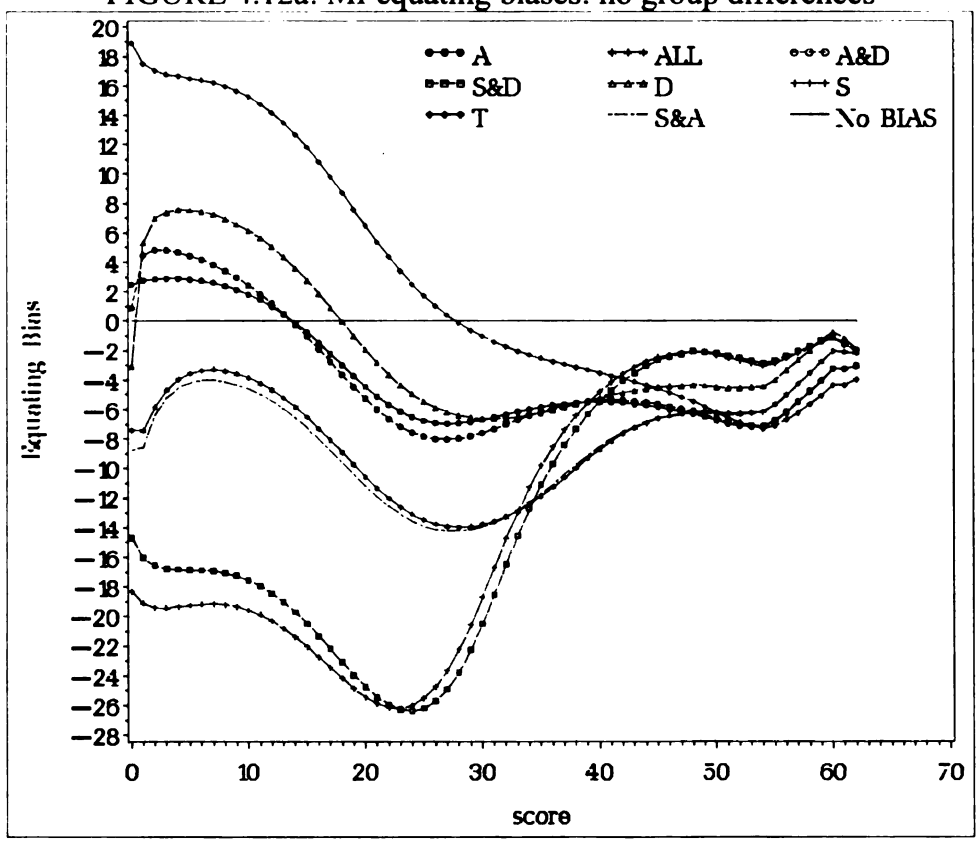

FIGURE 4.12b: MI equating biases: group differences

When groups differed in abilities, the equating biases displayed in Figure 4.12b diverged from zero, and different methods tended to have different equating biases. However, the methods that had relatively smaller equating biases included the PSE method that uses demographic variables (D), the PSE method that uses the combination of the anchor test and demographic variables (A&D) and the traditional PSE method (A). Even though these three methods had comparable equating biases, the D method seemed to be more preferable because it, on average, had the smallest equating biases. The traditional PSE method (A), the PSE method that uses sub-scores (S), and the method that uses the combination of sub-scores and demographic variables (S&D) had very large equating biases at the low end of the score scale. The traditional PSE method (T) had large positive biases, while the S and S&D methods had large negative biases at the low end of the score scale. PSE methods that use sub-scores (S) as the covariates to impute missing data had large negative biases.

CHAPTER 5

SUMMARY AND DISCUSSION

5.1 Background

This study investigated if it was feasible to use collateral information about examinees in the post-stratification equating (PSE) method to improve the quality of equating in terms of standard errors of equating and equating biases. Collateral information about examinees investigated in this study included five sub-scores, the anchor test score, and 24 examinees' demographic variables. The primary objective of this study was to explore the feasibility of using this information to reduce equating biases and standard errors of equating produced by the PSE method.

Theoretically, equating test scores using the PSE method, the equating method that uses the Non-Equivalent Groups with Anchor Test (NEAT) design, is achieved when score frequencies of missing data are obtained to construct synthetic population functions. For the NEAT design, there are two test forms (e.g., X and Y) and the anchor test (A), and there are two samples of examinees drawn from two different populations (e.g., P and Q) (von Davier, Holand, & Thayer, 2004). Examinees take each of these two test forms plus the anchor test. Therefore, examinees taking X will have missing scores on Y, while examinees taking Y have missing scores on X. The assumptions about the test score distribution conditional on the anchor test score have to be invoked and they are assumed to be invariant for both P and Q. By using these assumptions, score frequencies of missing data are obtained. Once score frequencies of missing data are obtained, synthetic population functions which are the mixture of observed score frequencies (from examinees' responses) and predicted score frequencies (from missing data) are

constructed in the form of $fx=wfx_P+(1-w)fx_Q$ and $fy=wfy_P+(1-w)fy_Q$, where f denotes the score frequency and w the weight given to P (Braun, & Holland, 1982). Then the PSE equating is carried out by applying fx and fy to the equipercentile function.

Researchers have shown that, when groups differ greatly in abilities, the PSE method produces larger equating biases than the chained equating. Therefore, the chained equating method has been used more widely than the PSE method, even though the PSE method has a stronger theoretical foundation than the chained equating method (Kolen, 1990). Specifically, the PSE method identifies the target population of the equating function defined by the synthetic population functions mentioned above, but the chained equating method does not define its target population. The equating function is defined for a "single population" and when groups differ greatly, the chained equating method has theoretical shortcomings in this regard because we do not know what group of examinees its equating function is computed for.

The large PSE equating biases are the result of the inaccurate prediction of missing data. In other words, the role of A as a conditional variable is not successful in obtaining the score frequencies of missing data. Wang and Brennan (2009) therefore replaced A with the "anchor test true score" in order to improve the equating biases of the PSE method. This method is called "the modified frequency estimation method" or "the modified PSE method" in this study.

This study explored an alternative strategy to improve equating biases by using more collateral information about examinees as a conditional variable. This is extremely necessary when groups differ greatly in abilities, only a single anchor test score and anchor test true score may not impute score frequencies of missing data precisely well.

Therefore, using more collateral information to increase the prediction of missing data is needed so that equating biases can be reduced. The second objective of this study was to investigate how precise the collateral information predicts score frequencies of missing data. Better prediction of missing data was expected to be associated with smaller equating biases.

Since investigations of collateral information in equating literature have not provided illuminating findings about its impact on standard errors of equating, this study also explored the effect of using collateral information about examinees on standard errors of equating of the PSE method.

When collateral information about examinees are available, this study proposed using two different ways to use this information to construct synthetic population functions needed for equating test scores under the PSE method. The first method is the propensity score method (Rosenbaum, & Robin, 1983) by which missing data occurring when equating data are collected through the NEAT design are predicted by invoking a conditional distribution of scores conditional on the propensity score. That is, the propensity score was used to replace the anchor test score in the traditional PSE method. The second method is the multiple imputation method (Rubin, 1987) by which missing data were imputed using available data without any assumptions.

Specifically, by using the propensity score method, the conditional distribution assumption of the traditional PSE method, which is "the conditional distribution of score given the anchor test score", were modified by replacing the anchor test score with the propensity score as a conditional variable. This modified assumption is interesting in that using more information about examinees plus the anchor test score to create a conditional

variable in the PSE method would be more realistic—it would be more held. However, when the multiple imputation (MI) method was used, it imputed missing data directly from observed collateral information. This was achieved without any changes in the PSE assumptions.

5.2 Study Design

This study used both simulation data and empirical data. For the simulation study, four factors investigated included two test lengths (60 items and 40 items), two ability differences (no differences vs. large group differences), two missing data treatments (the propensity score method and the multiple imputation method), and eight combinations of collateral information about examinees. The eight combinations of collateral information included

1. Anchor test sore only (A)
2. Anchor test score, sub-scores, and demographic variables (ALL)
3. Anchor test score and demographic variables (A&D)
4. Sub-scores and demographic variables (S&D)
5. Demographic variables only (D)
6. Sub-scores only (S)
7. Anchor test true score (T)
8. Sub-scores and the anchor test (S&A).

100 data sets were replicated for the PS method, but the first 20 data sets of the 100 data sets were replicated for the MI method. However, 5 data sets were imputed using the MI method for each of the 20 data sets; therefore, 100 analyses were also performed for the MI method.

The data analysis has two parts. Part I was to prepare data for equating, and Part II was the equating part. Part I involves the sub-score estimation and the propensity score estimation. Sub-scores were estimated using the Haberman (2008) method which is based on the classical test theory (CTT). This method was chosen to estimate sub-scores used in this study because it produced sub-scores comparable to those obtained through other complicated techniques (Sinharay & Haberman, 2008), but it is much easier to be carried out and less time consuming in computations. The propensity scores were estimated through the logistic regression model. Note that, within this part, the MI method involved the sub-score estimation only, but the propensity estimation is not needed for the MI method. Part II used the results from Part I as well as the anchor test score and the raw score to carry out equating.

For the real data analysis, the Mathematics Teaching for the 21st (MT21) Century was used. MT21 is a small international comparative study of mathematics teacher preparations of six countries. MT21 has two test forms each measuring five components of content knowledge—algebra, data, function, geometry, and number. This study manipulated the MT21 data such that the investigation of the effect of group differences in terms of abilities on the standard errors of equating and equating biases was feasible. Specifically, for the condition where there were no group differences, all cases in the MT21 data were used. However, for the condition where there were group differences, Taiwan and Korea data were excluded from the test form 1, and Bulgaria and Mexico data were excluded from the test form 2. Equating was performed using the KE software because it estimates standard errors of equating used for the comparisons between equating methods. Equating biases were defined as the differences between the obtained

equating function and the criterion equating function obtained through the IRT true score equating.

5.3 Results from the Simulation Data

5.3.1 The Prediction of Score Frequencies of Missing Data

One of the objectives of this study is to evaluate the accuracy of the PSE methods that use collateral information in predicting score frequencies of missing data. This evaluation was to find information to address why a certain PSE equating method had smaller or larger equating biases. Small equating biases are expected to be associated with accurate predictions of score frequencies of missing data.

Results from the simulation study showed different predictions of missing data between the propensity score method and the multiple imputation method. For the propensity score method, PSE methods that use sub-scores had more accurate predictions of score frequencies missing data than other methods when there were group differences in the long test condition. In a short test condition, these PSE methods had larger chi-square and LR chi-square statistics, meaning that they did not predict score frequencies of missing data well. However, where there were no group differences the PSE method that uses the combination of demographic variables and the anchor test score, and the PSE method that uses demographic variables had smaller chi-square statistics and LR chi-square statistics than other methods, indicating that these two methods when used with the propensity score method better predicted score frequencies of missing data. But when there were group differences, these two methods could not better predict missing data than the PSE methods that use sub-scores.

However, for the multiple imputation method, the PSE methods that use sub-scores as a component of predictors of missing data had smaller chi-square and LR chi-square statistics when there were no group differences. This was true for both long and short test. However, when there were group differences the PSE method that uses demographic variables had smallest chi-square and LR chi-square statistics.

5.3.2 Comparisons between the Traditional PSE Method, the Modified PSE Method, and the PSE Methods that Use Collateral Information

The second objective of this study is to assess the comparability of predicted score frequencies of missing among the proposed methods, the traditional PSE method and the modified PSE methods.

For the propensity score method of missing data treatment, when there were no group differences, the methods that had the smallest chi-square and LR chi-square statistics included the PSE method that uses the combination of sub-scores and demographic variables, and the method that uses demographic variables, followed by the modified PSE method and the traditional PSE method. When there were group differences in the long test condition, the methods that had smaller chi-square and LR chi-square statistics included ones that involved sub-scores. But when there were group differences in the short test condition, the modified PSE method and the traditional PSE method had smaller chi-square and LR chi-square statistics than other methods.

For the multiple imputation method of missing data treatment, when there were no group differences in both long and short test conditions, the methods that involve sub-scores had the smallest chi-square and LR chi-square statistics than other methods. But when there were group differences, the PSE method that use demographic variables had

smaller chi-square and LR chi-square statistics, followed by the modified PSE method, and the PSE method that uses all collateral information. This was true for both long and short tests.

5.3.3 Standard Errors of Equating and Equating Biases of the PSE Methods that Use Collateral Information

This study proposed using collateral information in the PSE equating method. The equating quality of these methods was evaluated by standard errors of equating and equating biases. The third objective of this study evaluated standard errors of equating and equating biases of the PSE methods that use collateral information. This evaluation assessed the quality of these PSE methods.

For the propensity score method of missing data treatment, all PSE methods that use collateral information had large standard errors of equating. Among these methods, the method that uses the combination of anchor test and demographic variables had smaller standard errors of equating, followed by the method that uses demographic variables. These two methods still had the smallest standard errors of equating in the short test condition.

In terms of equating biases, all PSE methods had comparable equating biases as seen in Figure 4.2a. When there were group differences, the methods that use sub-score had the smallest equating biases (Figure 4.2b). However, this was not true for the short test condition. In the short test condition, the methods that involved sub-scores had larger positive equating biases than other methods (see Figures 4.2c and 4.2d).

5.3.4 Comparisons between the PSE Methods that Use Collateral Information, the Traditional PSE Method, and the Modified PSE Method in Terms of Standard Errors of Equating and Equating Biases

The fourth objective of this study is to compare standard errors of equating and equating biases between the PSE methods that use collateral information, the traditional PSE method, and the modified PSE method.

For the propensity score method of missing data treatment, standard errors of equating of the traditional PSE method and the modified PSE method tended to be smaller than other methods in all conditions. The PSE method that use the combination of anchor test and demographic variables also had small standard errors of equating comparable to the traditional PSE method and the modified PSE method.

In terms of equating biases, the methods that involve sub-scores had the smallest equating biases when there were group differences in the long test condition. However, in the short test condition when there were group differences, the traditional PSE method and the modified PSE method had smaller equating biases as seen in Figure 4.2d. In addition, Figure 4.2c shows that the methods that involve sub-scores had large equating biases than other methods. These results show that sub-scores did not provide much value when tests had small number of items.

For the multiple imputation of missing data treatment, all methods had comparable small standard errors of equating as seen in Figures 4.5a-4.5d. In terms of equating biases, when there were no group differences, the traditional PSE method, the modified PSE method, and all PSE methods that use sub-scores had smaller equating biases than other methods (see Figures 4.6a and 4.6c). However, when there were group

differences, the modified PSE method and the PSE method that uses demographic variables had the smallest equating biases (see Figures 4.6b and 4.6d).

5.4. Results from the Empirical Data

This study used the Mathematics Teaching in the twenty first Century (MT21) data to investigate the impact of collateral information used in the post-stratification equating (PSE) method on standard errors of equating and equating biases. Data were manipulated such that the impact of group differences could be investigated.

For the propensity score method, standard errors of equating of all equating methods that used different sets of collateral information had similar standard errors equating with the maximum less than 1.00, when there were no group differences. Even though standard errors of equating were comparable, the traditional PSE method and the modified PSE method had the smallest standard errors of equating. However, when there were group differences, the modified PSE method produced stable standard errors of equating, no matter how large the group differences were. But other methods had even larger standard errors of equating.

In terms of equating biases, when there were no group differences, the PSE method that uses the combination of the anchor test and demographic variables and the method that uses demographic variables had smallest equating biases, followed by the traditional PSE method and the modified PSE method. However, when there were group differences, the PSE method that uses sub-scores became the method that, on average, had smallest equating biases as seen in Figure 4.10d. All other methods, except for the modified PSE method, that used different collateral information had even larger negative equating biases. The modified PSE method had large positive equating biases at the low

end of the score scale and large negative equating biases at the high end of the score scale. Even though the PSE method that uses sub-score had smallest equating biases, the other methods that use sub-scores as a component in the propensity score estimation had large equating biases.

For the multiple imputation method, all equating methods that use different sets of collateral information had comparable standard errors of equating, but when there were group differences, standard errors of equating of all methods were more different as seen in Figures 4.11a and 4.11b. No matter how large the group differences were, standard errors of equating of all methods were fairly small.

In terms of equating biases, similarly to the propensity score method, all equating methods had comparable equating biases when there were no group differences. The traditional PSE method, the method that uses the combination of the anchor test score and demographic variables, and the PSE method that uses demographic variables had relatively smaller equating biases than other methods. When groups differ greatly in abilities, nearly all methods tended to increase equating biases to more negative values. But the equating biases of the PSE method that uses sub-scores did not have much impact from group differences, and therefore, on average, it had smallest equating biases, even though equating biases were large.

5.5 Discussion

There are some important issues found in this study to be discussed in this section. These issues discussed below are related to the efficacy, and feasibility of using collateral information in the PSE equating methods, as well as the recommendation to use collateral information in practice.

5.5.1 Why Collateral Information Is Necessary? and Its Efficiency in the PSE Method

When equating data are collected through the Non-Equivalent Groups with Anchor Test (NEAT) design, two observed score equating method commonly used include the chain equipercentile equating method and the post-stratification equating method. However, the chain equipercentile equating method has been more widely used because it has smaller equating biases when groups of examinees differ greatly in abilities. The chain equating has been questioned and the question is centered on the target population of the equating function. More specifically, the equating function is computed for a single population. Typically, when the NEAT design is used to collect equating data, there are two groups of examinees each taking test X or test Y plus the anchor test (A). When these groups of examinees differ greatly in abilities, it implies that they are drawn from two distinct populations. Therefore, what is population the equating function is computed for is the question not addressed by the advocates of the chain equipercentile method.

Unlike the chain equipercentile method, the PSE method defines the target population of equating function but unfortunately has larger equating biases than the chain equipercentile method. Large equating biases are associated with the realistic of the conditional assumptions of the PSE method that are invoked to compute score frequencies of missing data, especially when group differences are greater. This is the major reason why practitioners have preferred the chain equipercentile method to the PSE method.

The traditional PSE method has two assumptions about missing data invoked to compute score frequencies of missing data. As mentioned in von Davier, Holland, and

Thayer (2004), "the conditional distribution of X given A is population invariant" and "the conditional distribution of Y given A is population invariant." These assumptions seem to be acceptably held when there are no group differences. But when groups differ greatly in abilities, the role of A as a conditional variable has been questioned (e.g., Wang & Brennan, 2009). When A fails to adjust for group differences, larger equating biases are expected.

There are research studies attempting to reduce equating biases when groups differ greatly in abilities. These studies ranged from using a specially designed A (Holland & Sinharay, 2008) and anchor test true scores (Wang & Brennan, 2009). Alternatively, this present study investigated if collateral information about examinees is capable of reducing equating biases.

When collateral information is used to replace the anchor test, they have to be combined into a single piece of information, and the propensity score method provides an efficient way to combine this information into a form of propensity scores. In this study, examinees' propensity scores (Z) were computed using the logistic regression model and the propensity score is the probability of being administered the new test form (Test form Y) given a set of collateral information treated as a set covariates. If the examinee A took the old test form and had the same propensity score as the examinee B who took the new test form, these two students are said to be equivalent in terms of covariates. Therefore, it is reasonable to use the propensity score to replace the anchor test score in the process of PSE equating.

Since this study used more information in the form of the propensity score as a conditional variable, the traditional PSE method is thus modified by replacing A with Z.

These modified assumption are "the conditional distribution of X given Z is population invariant" and "the conditional distribution of Y given Z is population invariant." These modified assumptions are realistic because two groups of examinees having equivalent propensity scores (or collateral information) are likely to have similar proficiency level; therefore, their score distributions are more likely to be invariant than when using the anchor test score only as a conditional variable. This belief is evaluated in this study and supported by the simulation study, showing that the PSE methods that use collateral information had more accurate predicted score frequencies of missing data and smaller equating biases than the traditional PSE method and the modified PSE method.

The simulation study shows that when there were no group differences, all equating method had comparable equating biases. However, the methods that use sub-scores as predictors in the propensity score estimation are better than the traditional PSE method and the modified PSE method in the long test condition when group differences are greater. Based on this result, the new direction to reduce equating biases of the PSE method would be centered on the sub-scores, even though further studies are needed to refine its potential. The question why the sub-score is better than the anchor test and the anchor test true score in terms of the equating bias reduction is discussed below.

Why sub-scores have smaller equating biases than the anchor test score and the anchor test true score when used with the propensity score method might be because this study used a more powerful measurement model called the diagnostic assessment model to measure examinees' proficiencies. Precise measures of examinees' proficient are more appropriate to be used as a conditional variable to compute score frequencies of missing data. Sub-score estimation has been a new growing area in the psychometric field.

Psychometricians have recently increased attention to sub-score estimation (e.g., Roussos, Templin, & Henson, 2007; Haberman & Sinharya, 2008) because sub-scores provide more information about students' strengths and weaknesses than a single score. The Haberman's method of sub-score estimation used in this study estimates sub-scores using the classical test theory. This method is useful because it not only provides estimates of sub-scores comparable to those from other methods (Haberman & Sinharay, 2008) but it also is easier and less time consuming in computations. When sub-scores are used in the PSE method, they are expected to be more precise estimates of examinees' ability than the anchor test score which is a single score. They also better represent examinees' proficiency than the anchors test true score because they reflect complete dimensions of proficiency of examinees, similar to the multidimensional item response theory model (e.g., Reckase, 1997). However, the anchor test true score is just a combination of many dimensions of proficiency. Therefore, they are more realistic variables for predicting examinees' test score on the test form that they have not been administered to.

5.5.2 Propensity Score Method vs. Multiple Imputation Method

As indicated in the previous section, collateral information provides an alternative approach to reduce equating biases. When collateral information is obtained, we need a method to use this information for computing score frequencies of missing data required to construct the synthetic population functions which are the inputs for the PSE method. This study found that when there were no group differences, the propensity score and the multiple imputation methods tended to have comparable equating biases and predictions of score frequencies. However, they seemed to yield different results when there were

group differences in terms of abilities. Specifically, the propensity score method resulted in smaller equating biases for the PSE methods that use sub-scores or the combination of sub-scores and other collateral information about examinees. However, the demographic variables were most useful to predict score frequencies of missing data using the multiple imputation method. These results were consistent between simulation data and empirical data analyses.

In terms of standard errors of equating, the multiple imputation method had smaller values than the propensity score method. However, practitioner should be cautious when comparing the propensity score method with the multiple imputation method in terms of standard errors of equating because the standard errors of equating for the two methods were computed in this study using different approaches. The PS method estimated standard errors of equating using the Kernel Equating software in which standard errors of equating were estimated by the Tayler's series expansion method (von Davier, Holland, and Thayer, 2004), while the MI method uses the method similar to resampling methods.

5.5.3 Issue of the Length of Sub-scores

This study found that sub-scores for the long test had a potential to reduce equating biases, but this is not the case for the short test. This has an implication that equating biases can be reduced when long sub-scores are used with the propensity score method. This is reasonable in that sub-scores are precisely estimated in a long test because a long test is more reliable than the shorter test. This was consistent with Sinharay, Haberman, and Puhan (2007) where they noted that subscores are more meaningful for the long test.

5.5.4. Selection of Demographic Variables

This study shows that demographic variables have potential to reduce equating biases for the multiple imputation method. In the equating literature, the documentation about what demographic variables are useful for equating has been unclear. This study used demographic variables that are related to examinees' opportunity to learn in mathematics. These variables are highly correlated with test performance and thus they are expected to reduce equating biases because they can predict test scores of the missing data reasonably well.

When various demographic variables are available, practitioners should be cautious that not every set of demographic variables can reduce equating biases. Some demographic variables that are less relevant to examinees' proficiency are intuitively not associated with small equating biases. For example, Paek, Liu, and Oh (2008) found that using sex, ethnicity, and grade of students did not add much value to improve equating results.

5.5.6 Sub-score vs. Demographic Information

This study found that both sub-scores and demographic variables are feasible to reduce equating biases. The sub-score can reduce equating biases when the propensity score method was used in the situation when groups differ greatly. However, demographic variables are useful when they were used with the multiple imputation method to impute score frequencies of missing data. This study recommends that the sub-score is a better choice than the demographic variables because the sub-score is within the test. However, additional money is needed when one wants to gather good demographic information for the PSE equating.

5.5.7 Unfair Advantages Due to Equating Biases

The propensity score methods tend to yield positive biases at the low end of the score scale, indicating that low-achieving examinees will get an unfair advantage. Specifically, all propensity score equating methods overestimated equated scores, especially at the low end of the score scale (0-10 for the short test condition and 0-18 for the long test condition). For example, the sub-score and demographic variables gives an unfair advantage to examinees whose scores on the old form (test form 1) are less than 18 over those examinees performing better on the old form (test form 1). However, the minimum for the simulation data of this study is 30 for the long test condition and 18 for the short test condition, meaning that there are no examinees having scores in the range of 0-18 and 0-29 for the short test and long test, respectively. This pattern is called "zero frequencies." Therefore, the unfair advantage issue is not a major concern for the simulation data of this study. This is also true for the multiple imputation method.

But the propensity score methods that use sub-scores had large positive equating biases in the middle of the score scale. This indicates that using sub-scores with the propensity score method will give an unfair advantage to examinees whose abilities are about average, while other examinees will get a disadvantage.

For the multiple imputation method, when there were group differences, all methods had different patterns of biases. What groups get a more unfair advantage depends on the equating method used. Generally, low and high achieving examinees have either an unfair advantage or an unfair disadvantage, depending on the equating method used. However, as mentioned earlier, there are zero frequencies at the low and high end

of the score scale. Therefore, the equating bias is not a major problem that creates unfair (dis)advantages for this study.

5.5.8 A Potential for Bias When Rounding in Multiple Imputation

This study found that standard errors of equating and equating biases were poorly estimated at the bottom end of the scale. Rounding in multiple imputation method has been thought to be a cause of biases (Horton, Lipsitz, & Parzen, 2003). But it was found in this study that rounding the imputed scores to the discrete scores worked well at the low and high ends of the score scale. Specifically, the differences between imputed score frequencies and true frequencies are close to zero at the low and high ends of the score scale. Therefore, rounding is not a cause of poor estimates of equating biases and standard errors of equating. However, zero frequencies and the linear interpolation might be two sources of poor estimates of these statistics.

5.5.9 Effect of the Order of Group Differences between Examinees of P and Q

In the situation where there were group differences between populations of P and Q, this study simulated data such that the examinees of Q are more proficient than the examinees of P. This scenario is consistent with the practice of some testing programs such as the program that uses computerized testing. That is, the new test form (Y) administered to a new group of examinees is usually easier than the old form (X) because of the practice effect and cheating.

When the examinees of Q were more proficient than those of P, the propensity score methods and multiple imputation equating methods tended to result in positive equating biases, a finding is consistent with Holland and Sinharay (2007). However, this study did not investigate the situation where examinees of P are more proficient than

examinees of Q. For this condition, Holland and Sinharay (2007) found that equating biases were negative when P performs better than Q.

5.5.10 Requirements of Equating

The term "linking" refers to any function or transformation used to connect the scores on one test to those of another test. But "equating" is a special case of linking. A linking between scores on two tests to be considered an equating has to satisfy the following requirements (Dorans & Holland, 2000; Petersen, 2008):

- Same construct: The two tests must both be measures of the same characteristics (ability or skill).
- Equal reliability: Scores on the two forms are equally reliable.
- Symmetry: The transformation is invertible.
- Equity: It does not matter to examinees which test they take.
- Population invariance: The transformation is the same regardless of the groups from which it is derived.

Any score transformation that satisfies the above five requirements is considered equating (Dorans & Holland, 2000). This study did not check all of these requirements. The simulation data and empirical data used in this study seem to satisfy the first and second requirements. Moreover, since this study used the equipercentile equating function to derive comparable scores, the third requirement is thought to be satisfied by the definition of equipercentile equating function (Kolen, & Brennan, 2004). However, whether or not linking methods that use the propensity equating method and the multiple imputation method satisfy the remaining requirements is an interesting area for future research.

5.6 Implications

The results from this study provided some strategies that provide smaller equating biases than the PSE methods that have been used in practice: the traditional PSE method (e.g., von Davier, Holland, & Thayer, 2004) and the modified PSE method (Wang & Brennan, 2009). The findings from this study therefore have tremendous implications for practitioners. That is, this study shows that it is reasonable to use more collateral information to reduce equating biases when groups differ greatly in abilities. Two choices are provided in this study. The first choice is to use the propensity score method to combine sub-scores or the combination of sub-scores and other demographic variables. This choice works best with the propensity score method in the long test. The second choice is to use demographic variables with the multiple imputation method to predict score frequency of missing data directly without making conditional assumptions.

Sub-scores are more appealing than demographic variables because they are within the test; one does not have to spend times, costs, and energy to collect them for equating. This recommendation is proper for the equating purpose only but does not to say that demographic variables are not useful and should not be collected at all.

5.7 Limitations

There are some limitations of this study. First, this study used very special simulation data in which item responses were simulated but demographic variables from real data were fixed. Therefore, the sample size could not be varied because, by using demographic variables from the real data, 2627 rows of demographic variables from the empirical study had to be merged with the simulated item responses and thus every simulated data has 2627 cases. Second, for the empirical data analysis, the equating

biases were defined as the differences between the obtained equating function and the equating function from the IRT true scores. The simultaneous calibration was performed using the software BILOG-MG to obtain the estimates of item and abilities parameters for computing IRT true scores. However, MT21 data is thought to be multidimensional data because it measures multiple content areas, and therefore, item and ability parameters may be biased due to the multidimensionality issue.

5.8 Future Direction

There are many issues to be further studied to refine the results of the PSE methods that use collateral information.

First, it is obvious that equating biases and standard errors of equating at the low end of the score scale are large. These large equating biases and standard errors of equating were not precisely estimated because there were few examinees having low scores and because this study used the linear interpolation method to connect the score distribution at the low and high end of the score scale where score frequencies were sparse. Further study may continuize score distributions using the continuization method (von Davier, Holland, & Thayer, 2004) to solve the discreteness nature of score distribution and thus equating biases and standard errors of equating can be more precisely estimated.

Second, this study found in the empirical data analysis that the PSE method that uses sub-scores did not work in terms of equating biases. However, the result from simulation data analyses indicated that it outperformed other equating methods in the long test condition but it did not work in the short test condition. So the result from empirical data analysis is consistent to that of a simulation data analysis in the short test

condition. As noted in chapter 3, the MT21 data has some sub-sections that have very small number of items (e.g., the algebra section of Test form 1). Therefore, it is necessary to use another empirical data set that has lengthy subsections to see if the methods that involve sub-score really can reduce equating biases in practice.

Third, it is interesting to apply the kernel equating framework (von Davier, Holland, & Thayer, 2004) to the methods of this study because the kernel equating framework provides an efficient way to compute standard errors of equating and will improve the accuracy of standard errors of equating.

Fourth, this study shows that some PSE methods that use collateral information reduced equating biases (e.g., the method that uses sub-score), and in the longer test condition it had smaller equating biases than the traditional PSE method. As mentioned earlier, several previous studies highlighted that the traditional PSE method always produces larger equating biases than the chain equating when groups differ greatly in abilities. Hence, the chain equating has been used more widely in practice, although it has a theoretical shortcoming. Therefore, it is also interesting to compare the methods that use collateral information with the chained equipercentile method in terms of equating biases. This comparison will provide evidence for practitioners to determine if the PSE method with collateral information can become a better choice of equating methods as it is theoretically supposed to be than the chain equating method.

Fifth, it is interesting to apply the propensity score method to adjust for group differences in other equating designs in order to fully gain insight into its usefulness.

Finally, given that in some conditions the propensity score is more realistic to be a conditional variable to estimate score frequencies of missing data, it is possible to use

propensity scores to derive the classification consistency indices of two equated forms (Yi, Kim, & Brennan, 2007). This is usual when a test developer is interested in the extent to which an examinee who happens to take a particular form would have a consistent classification decision if he or she had taken an equated alternate form.

APPENDICES

Appendix A. FT Residuals for the Propensity Score Method

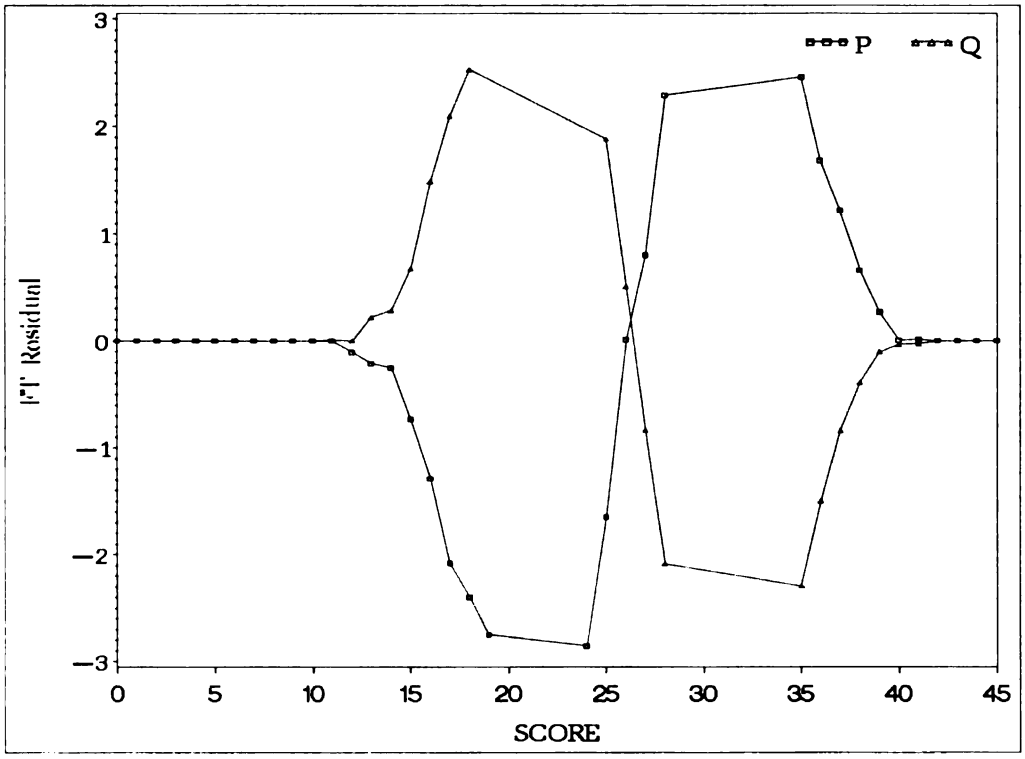

Figure A.1 No Group Differences, 45 Missing Items, Anchor Test Score (A)

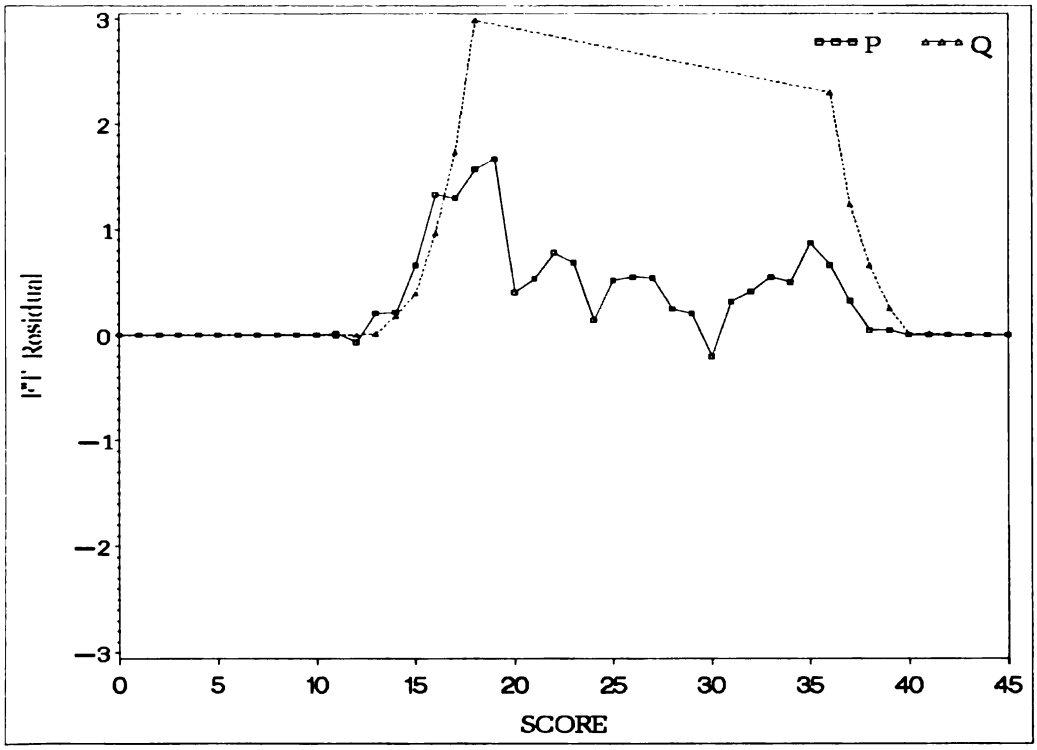

Figure A.2 No Group Differences, 45 Missing Items, All Collateral Information (ALL)

App

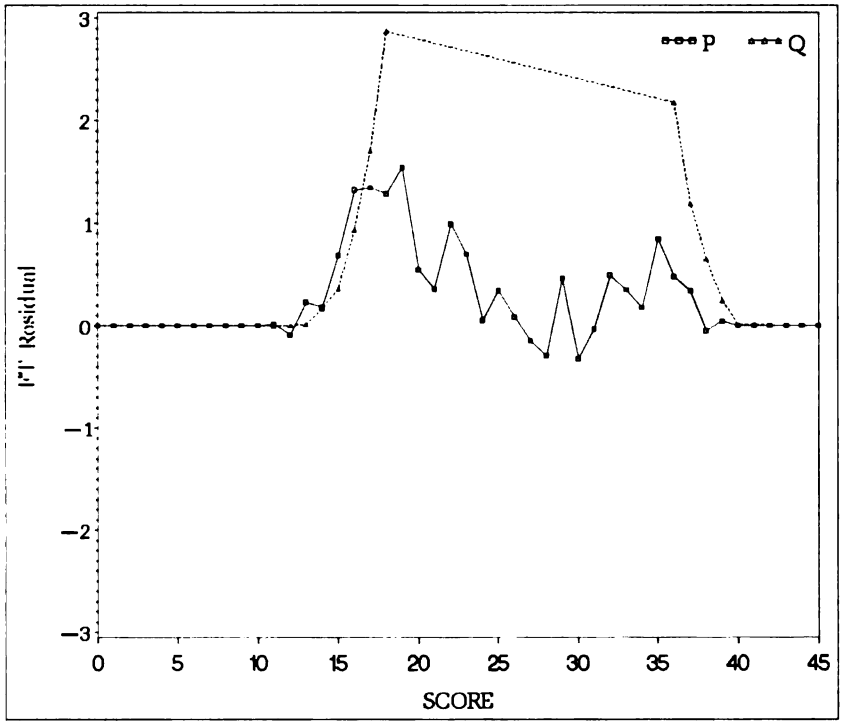

Figure A.3 No Group Differences, 45 Missing Items, Anchor Test Score and Demographic Variables (A&D)

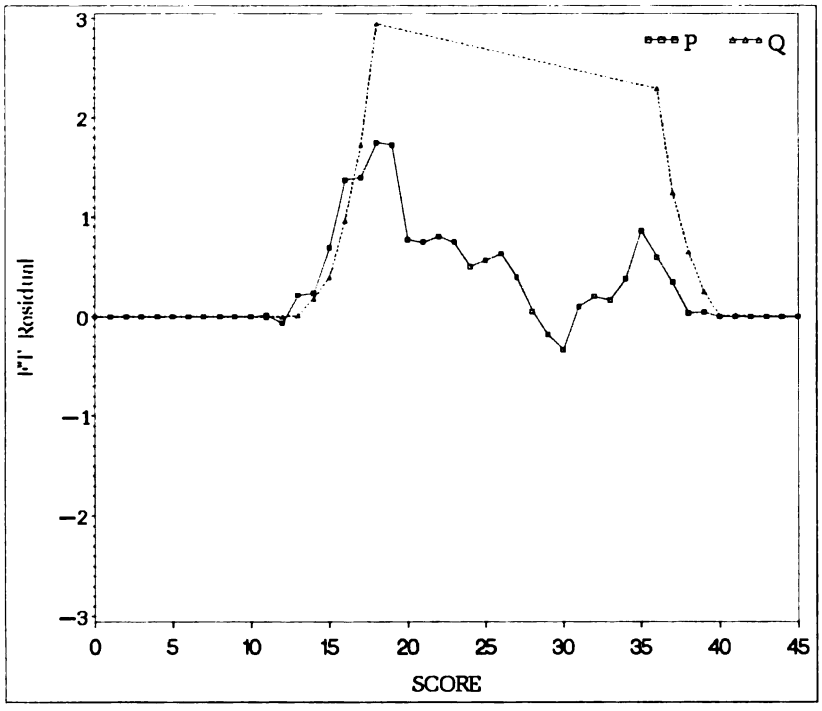

Figure A.4 No Group Differences, 45 Missing Items, Subscore and Demographic Variables (S&D)

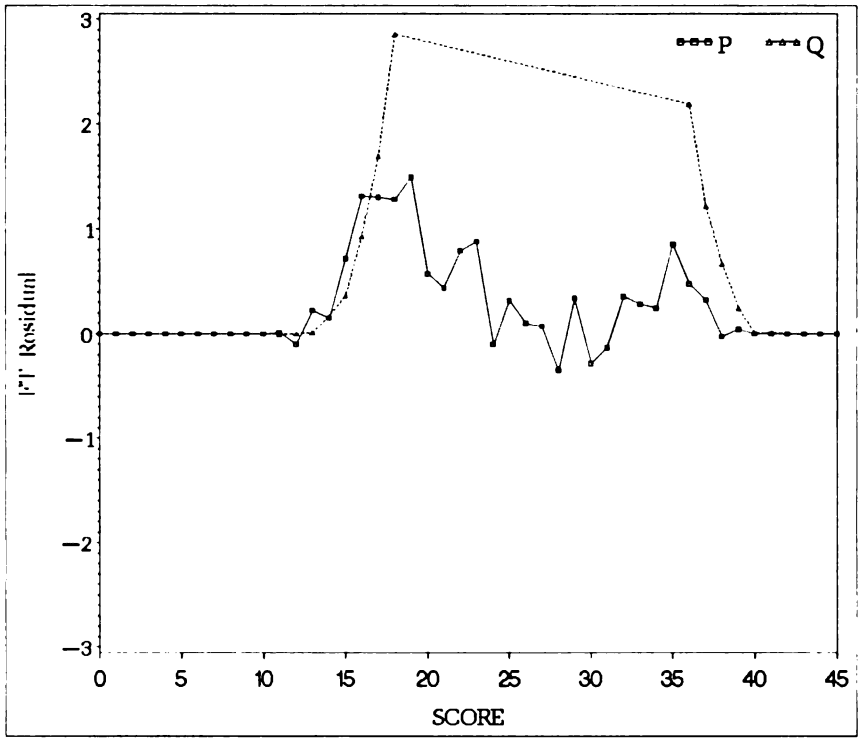

Figure A.5 No Group Differences, 45 Missing Items, Demographic Variables (D)

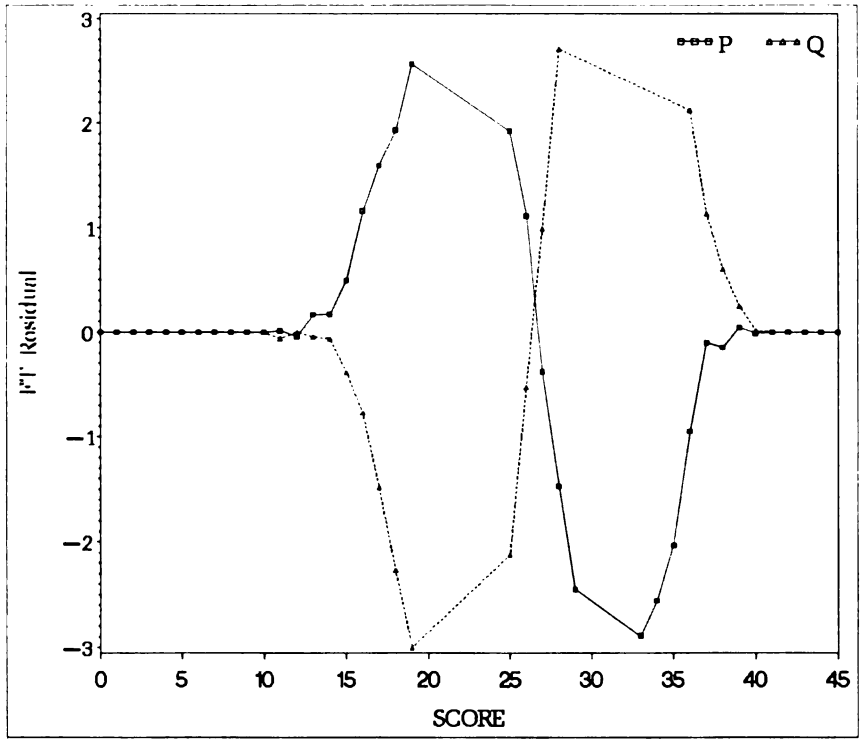

Figure A.6 No Group Differences, 45 Missing Items, Subscores (S)

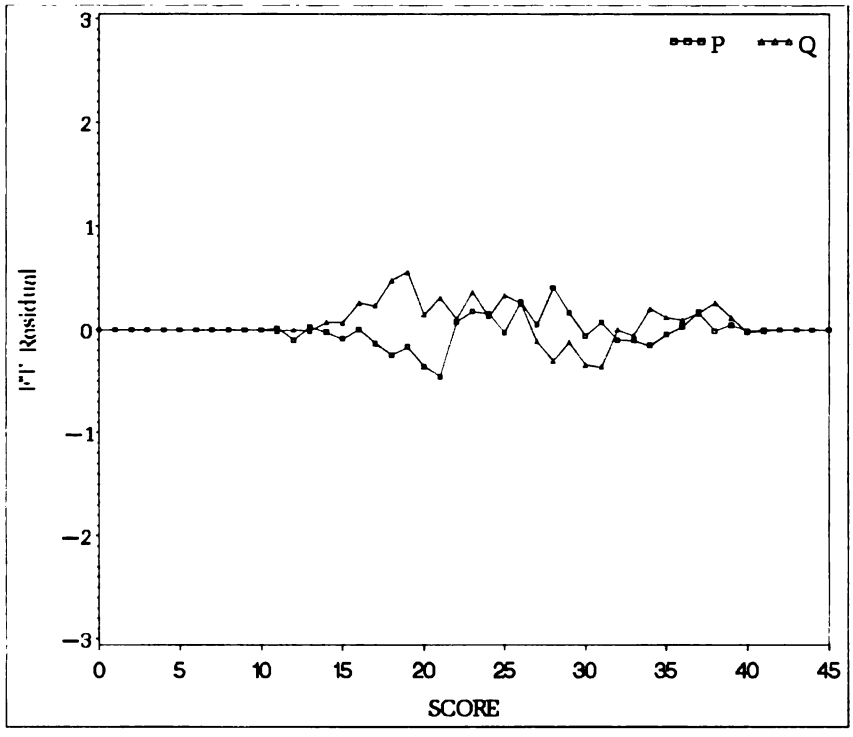

Figure A.7 No Group Differences, 45 Missing Items, Anchor Test True Score (T)

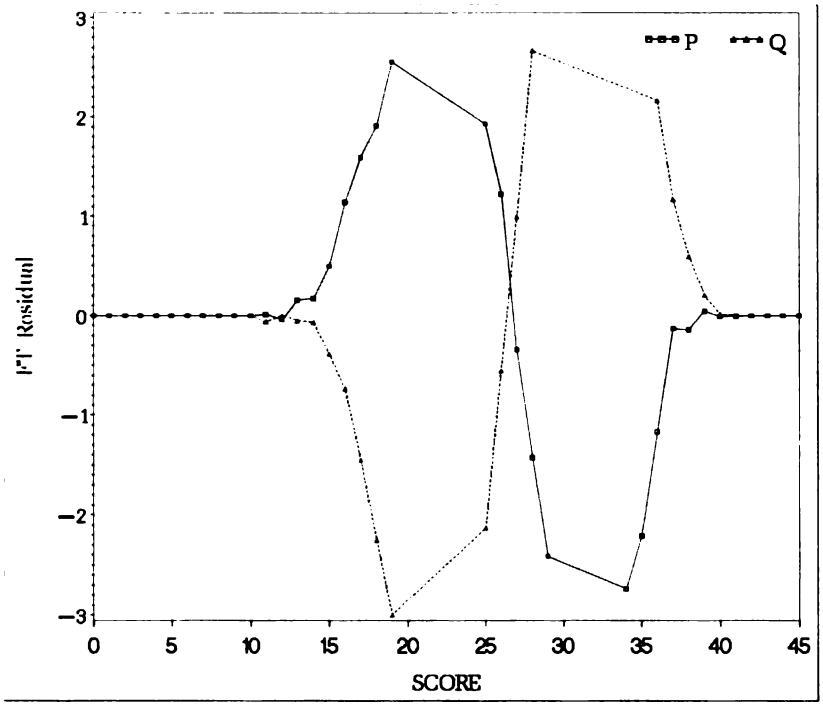

Figure A.8 No Group Differences, 45 Missing Items, Subscores and Anchor Test Score (S&A)

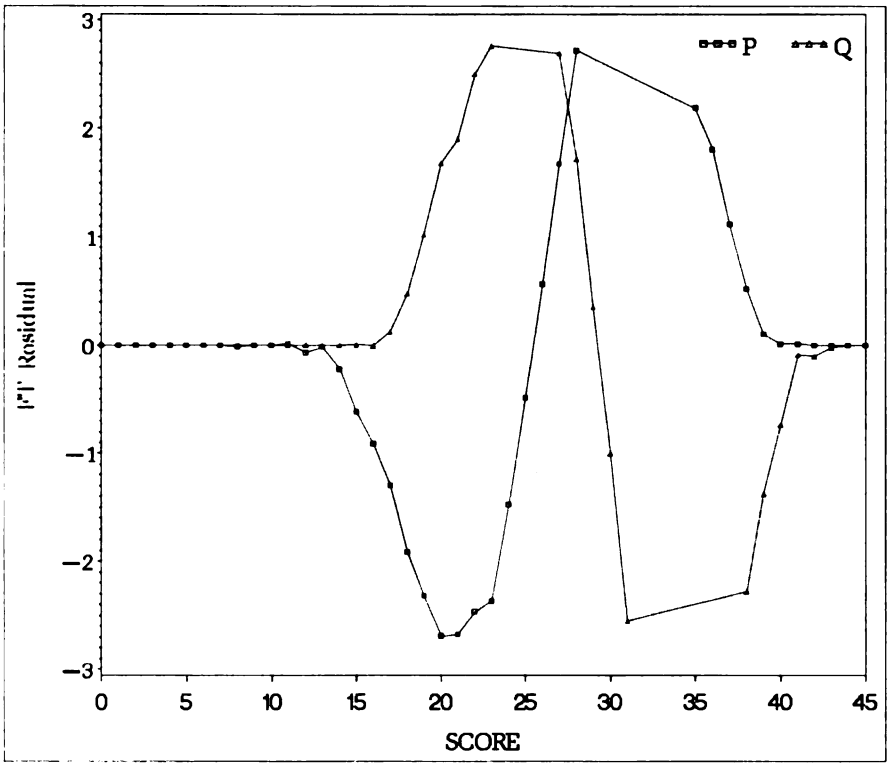

Figure A.9 Group Differences, 45 Missing Items, Anchor Test Score (A)

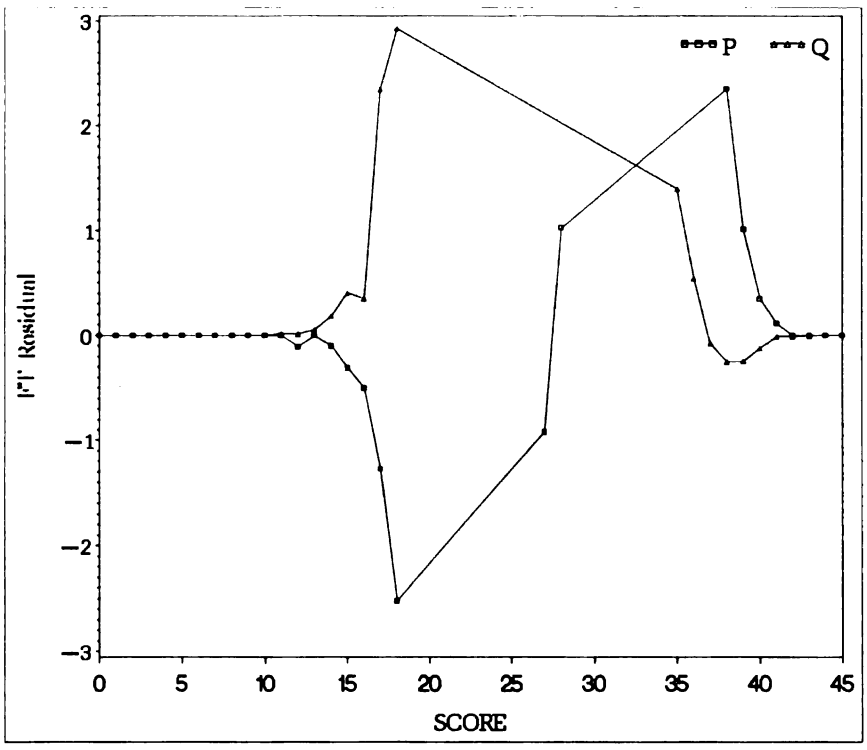

Figure A.10 Group Differences, 45 Missing Items, All Collateral Information (ALL)

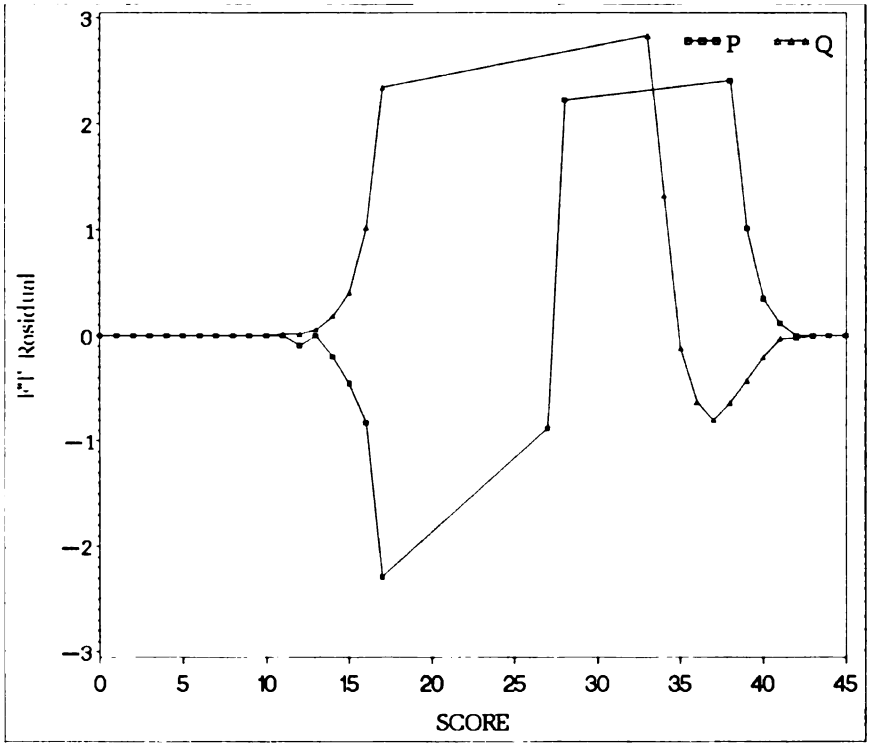

Figure A.11 Group Differences, 45 Missing Items, Anchor Test Score & Demographic variables (A&D)

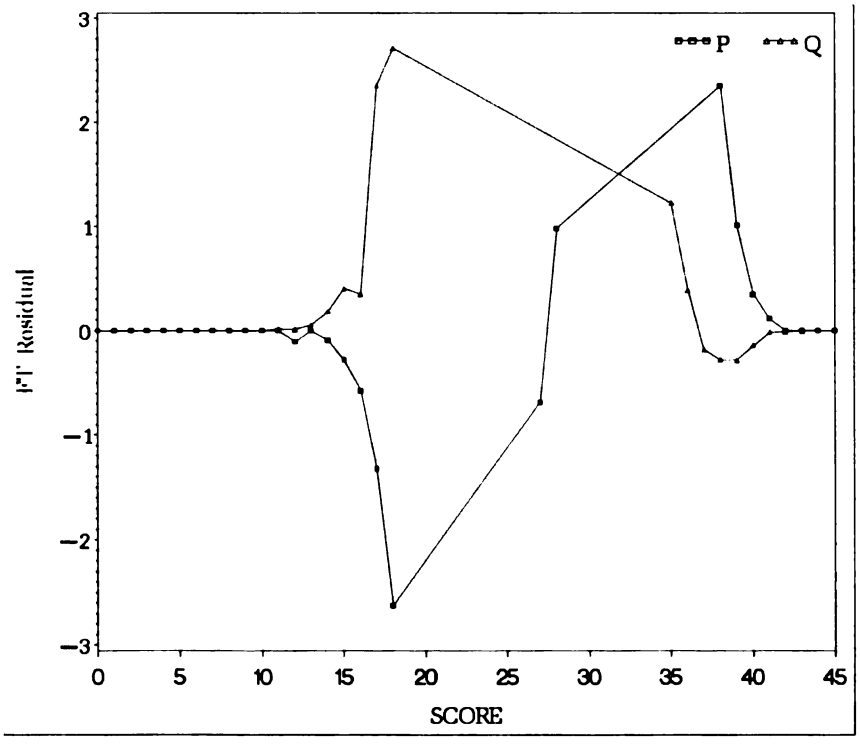

Figure A.12 Group Differences, 45 Missing Items, Subscores & Demographic variables (S&D)

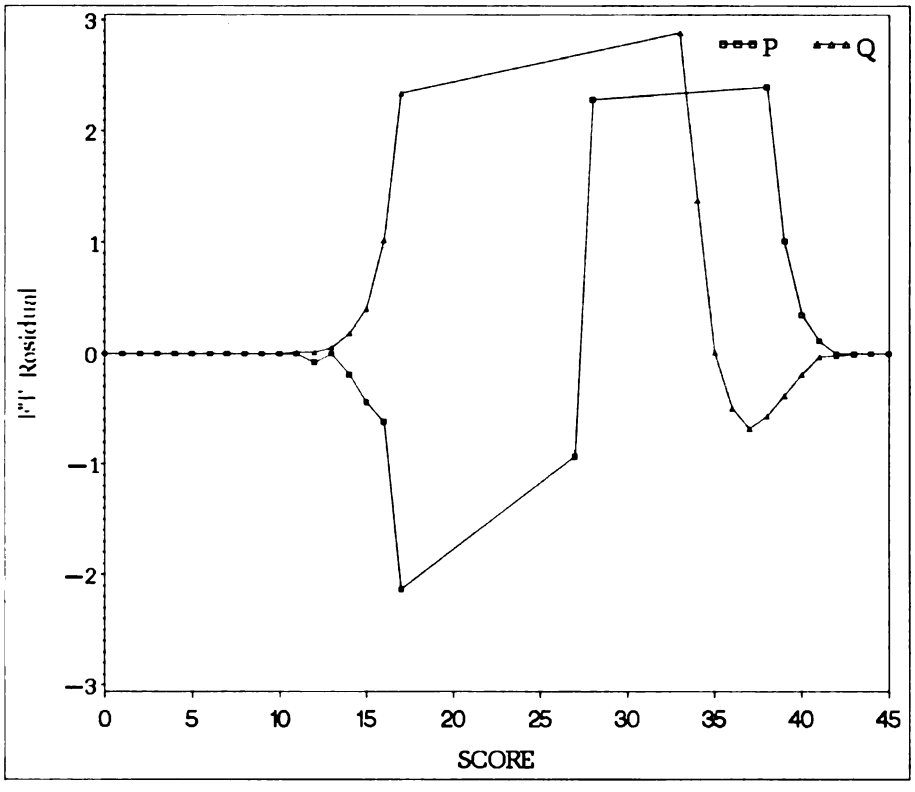

Figure A.13 Group Differences, 45 Missing Items, Demographic variables (D)

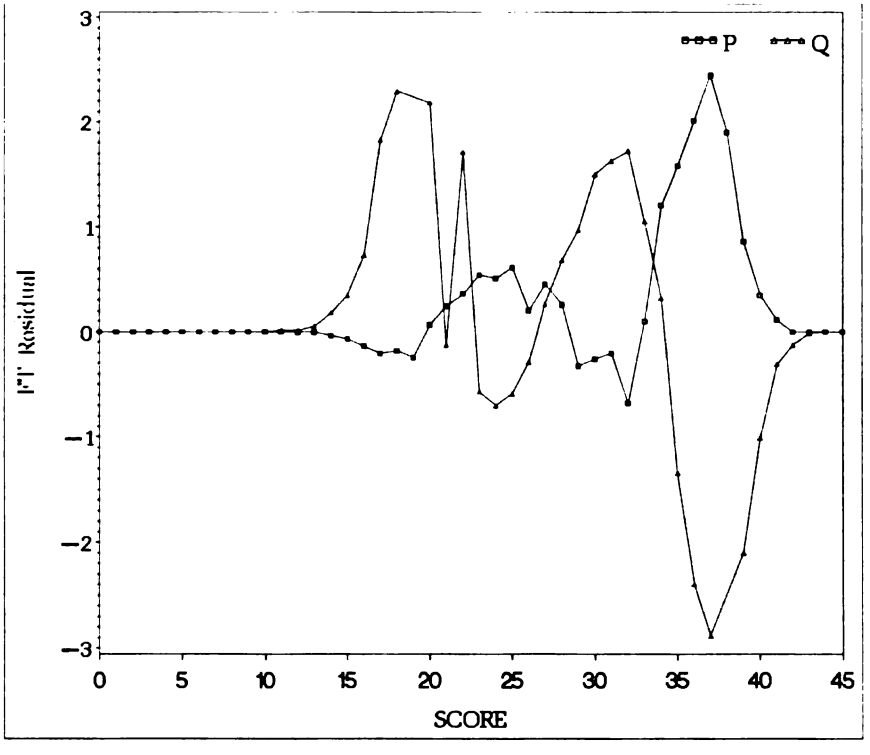

Figure A.14 Group Differences, 45 Missing Items, Subscores (S)

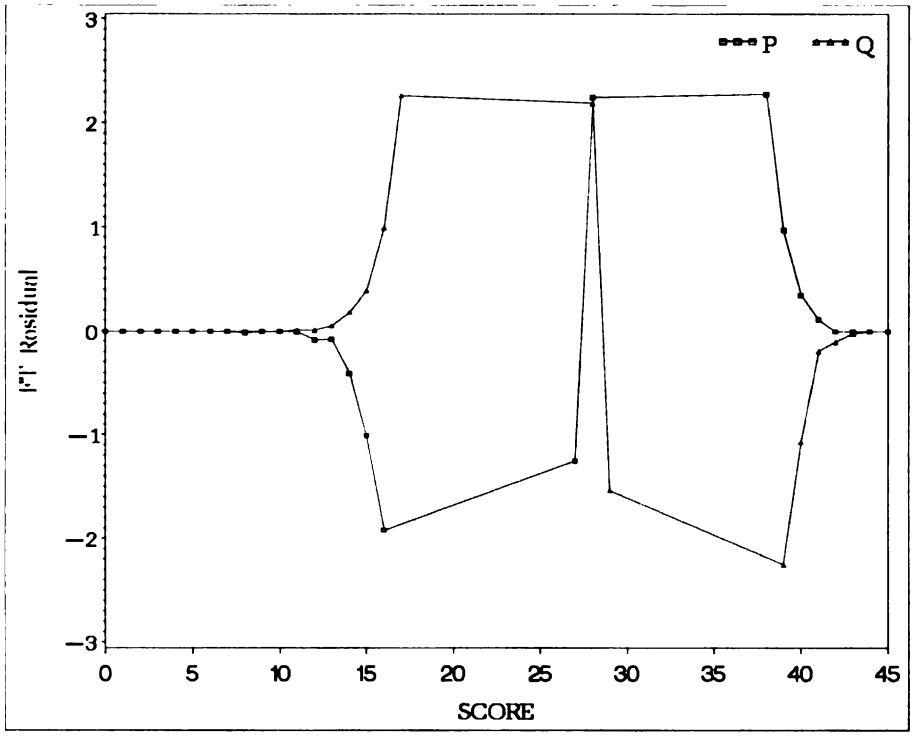

Figure A.15 Group Differences, 45 Missing Items, Anchor Test True Score (T)

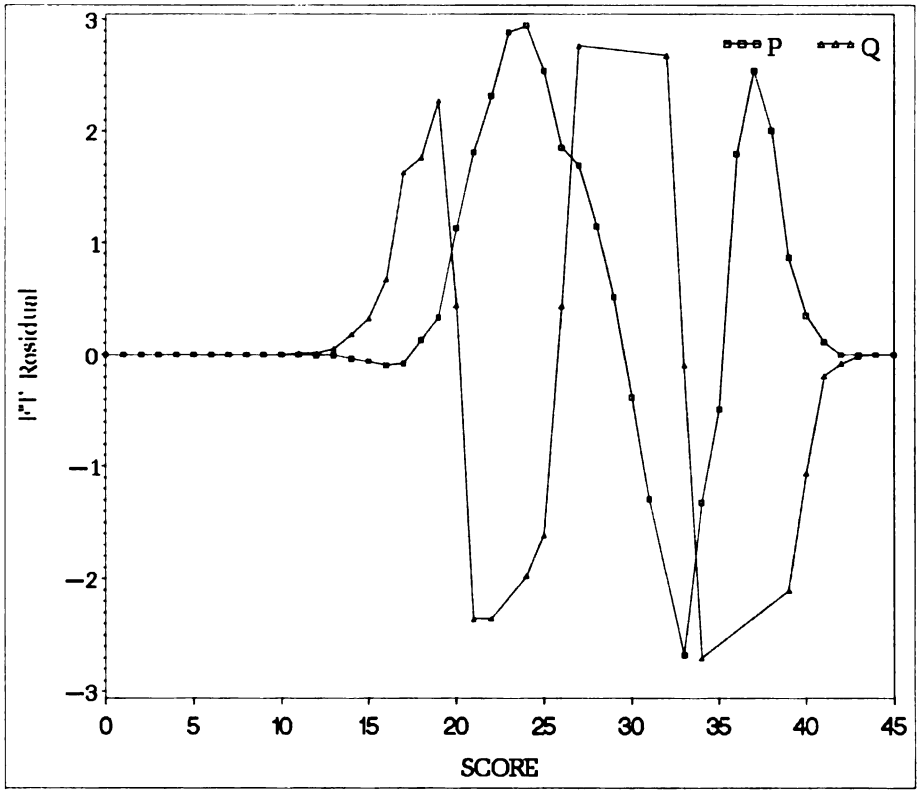

Figure A.16 Group Differences, 45 Missing Items, Subscores and Anchor Test Score (S&A)

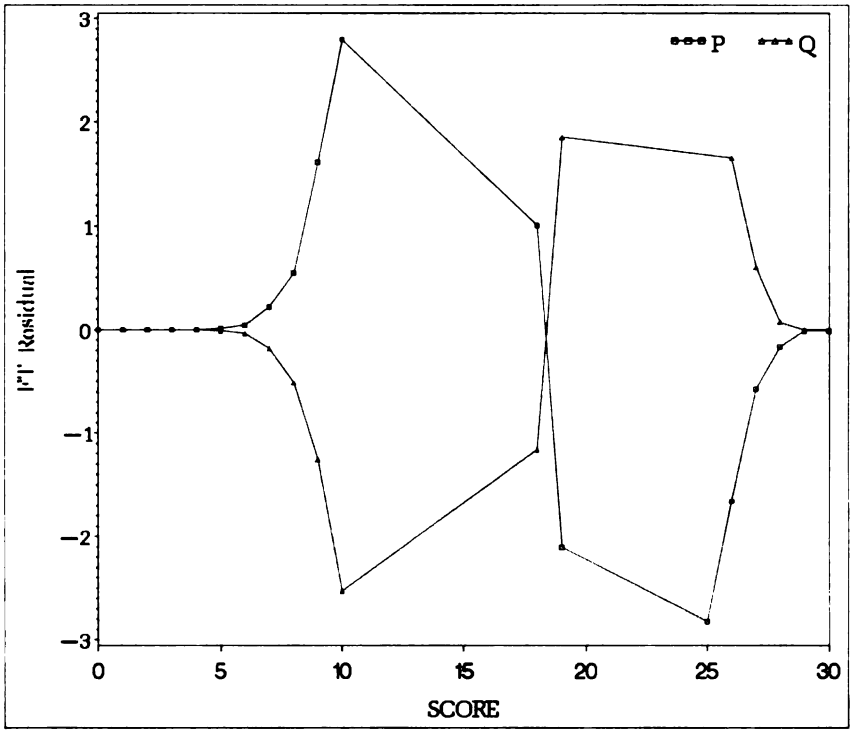

Figure A.17 No Group Differences, 30 Missing Items, Anchor Test Score (A)

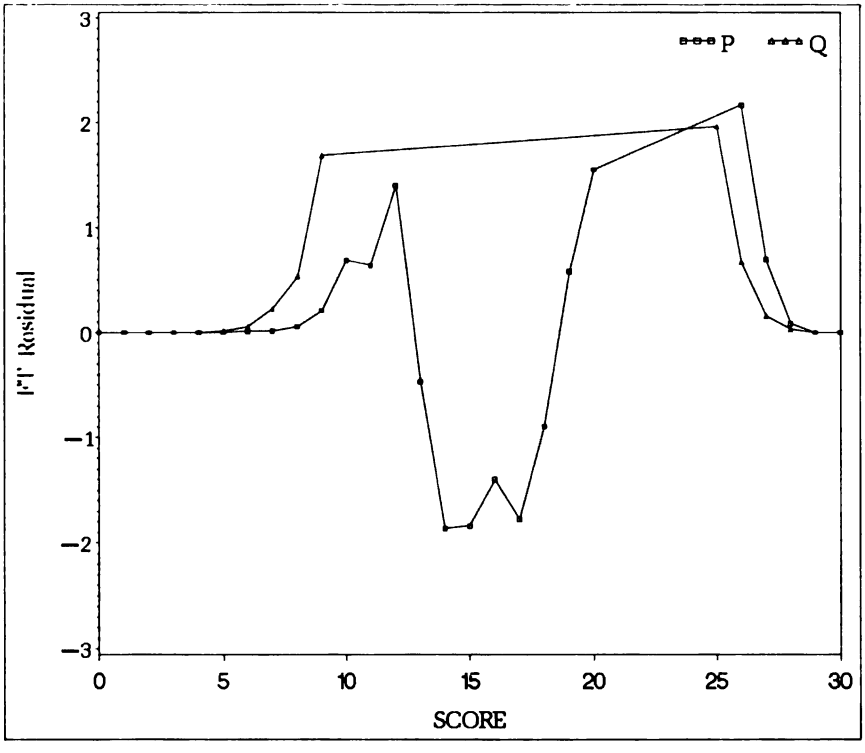

Figure A.18 No Group Differences, 30 Missing Items, All Collateral Information (ALL)

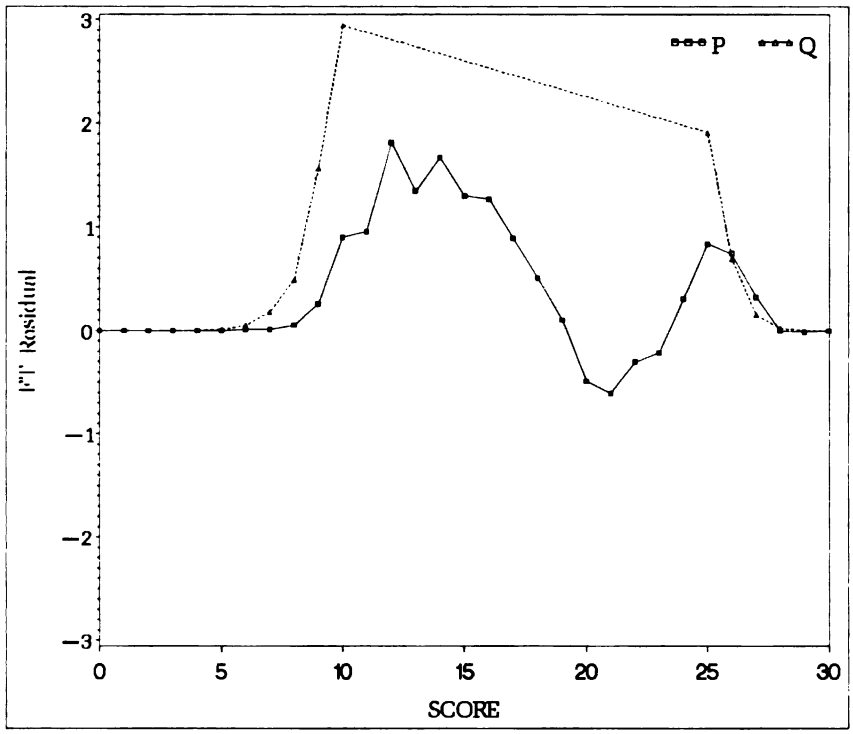

Figure A.19 No Group Differences, 30 Missing Items, Anchor Test Score and Demographic Variables (A&D)

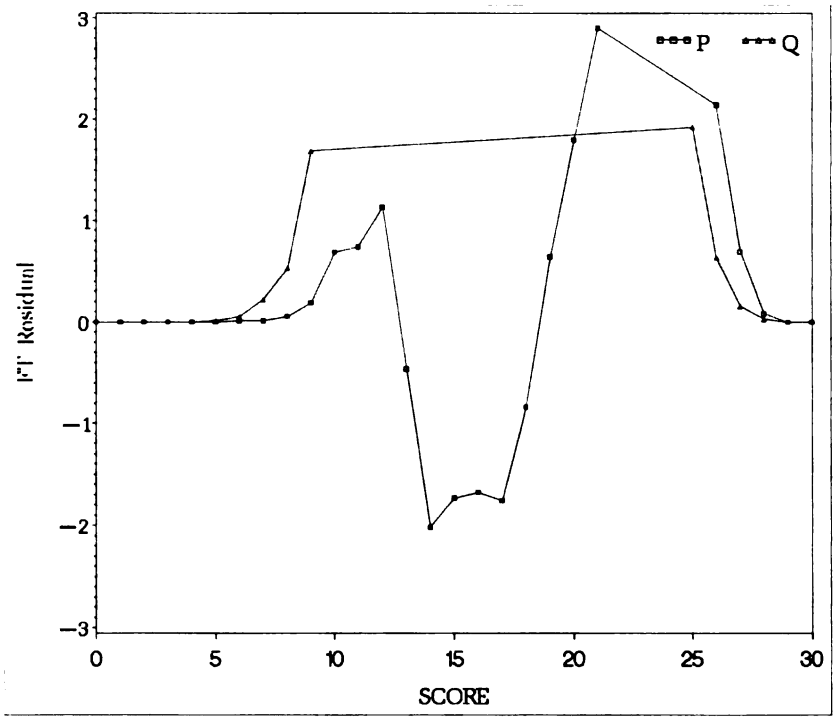

Figure A.20 No Group Differences, 30 Missing Items, Subscores and Demographic Variables (S&D)

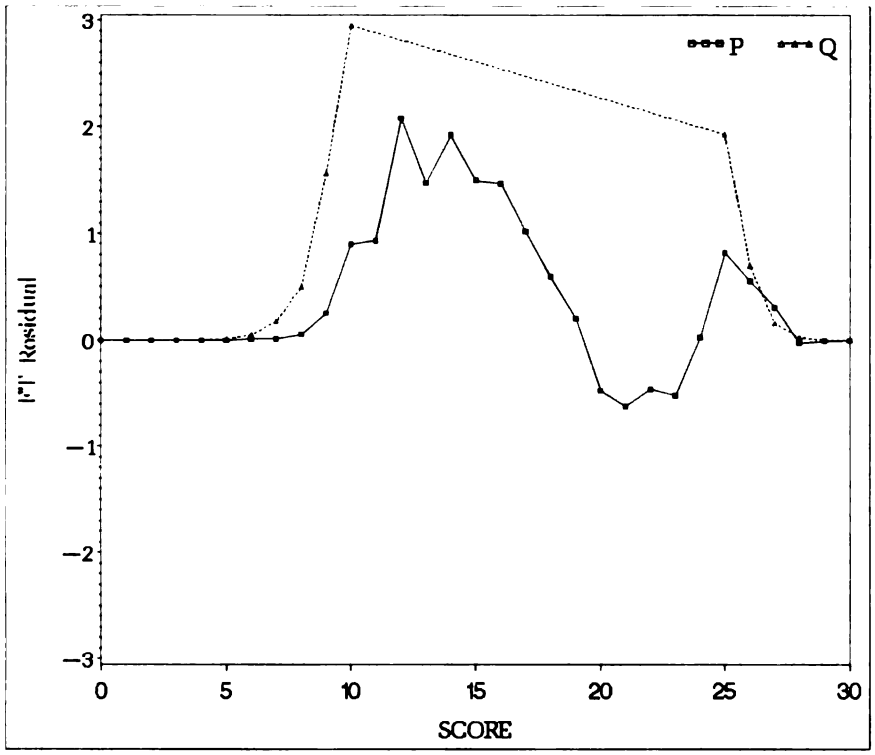

Figure A.21 No Group Differences, 30 Missing Items, Demographic Variables (D)

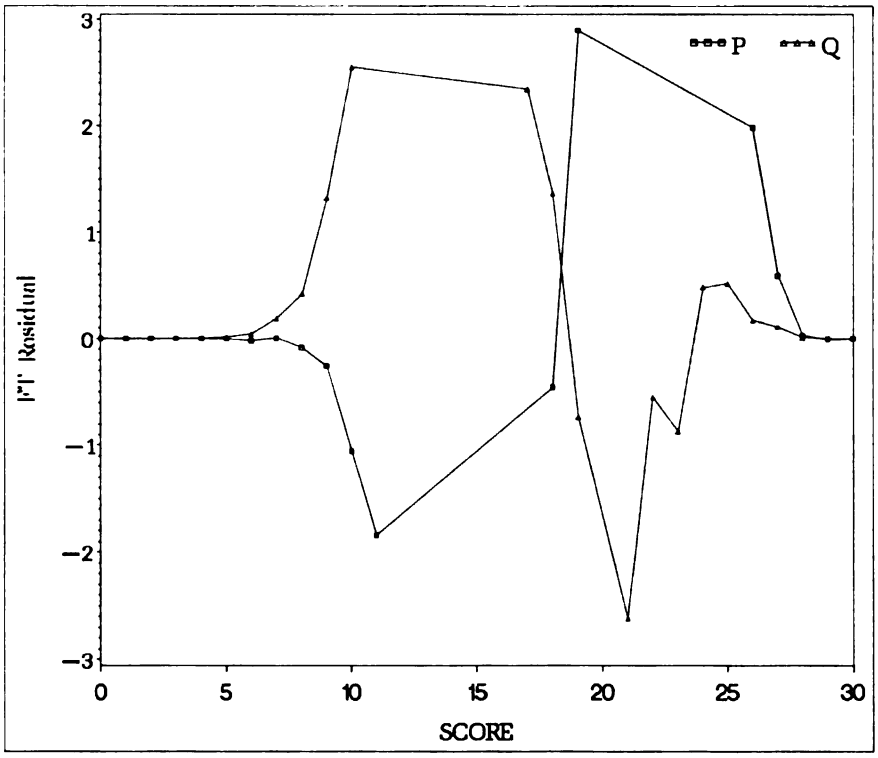

Figure A.22 No Group Differences, 30 Missing Items, Subscores (S)

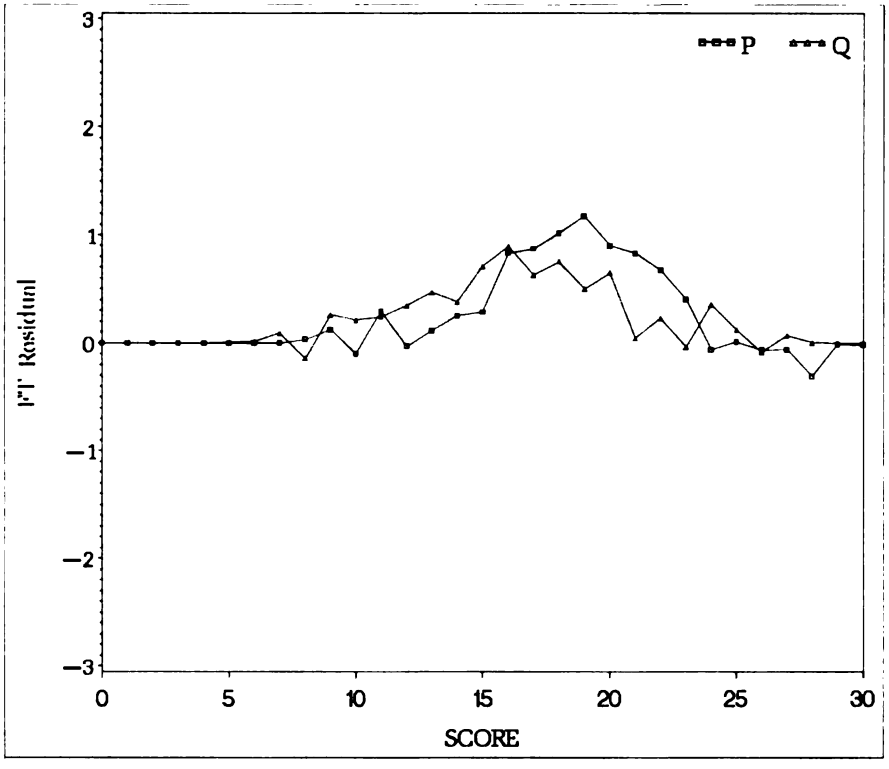

Figure A.23 No Group Differences, 30 Missing Items, Anchor Test True Score (T)

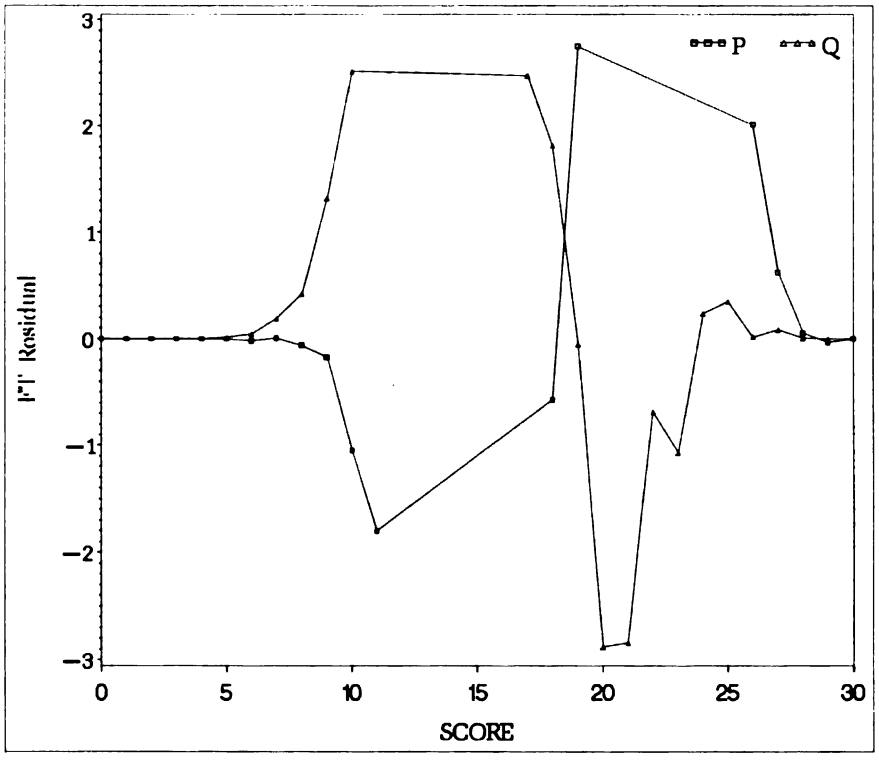

Figure A.24 Group Differences, 30 Missing Items, Subscores and Anchor Test Score (S&A)

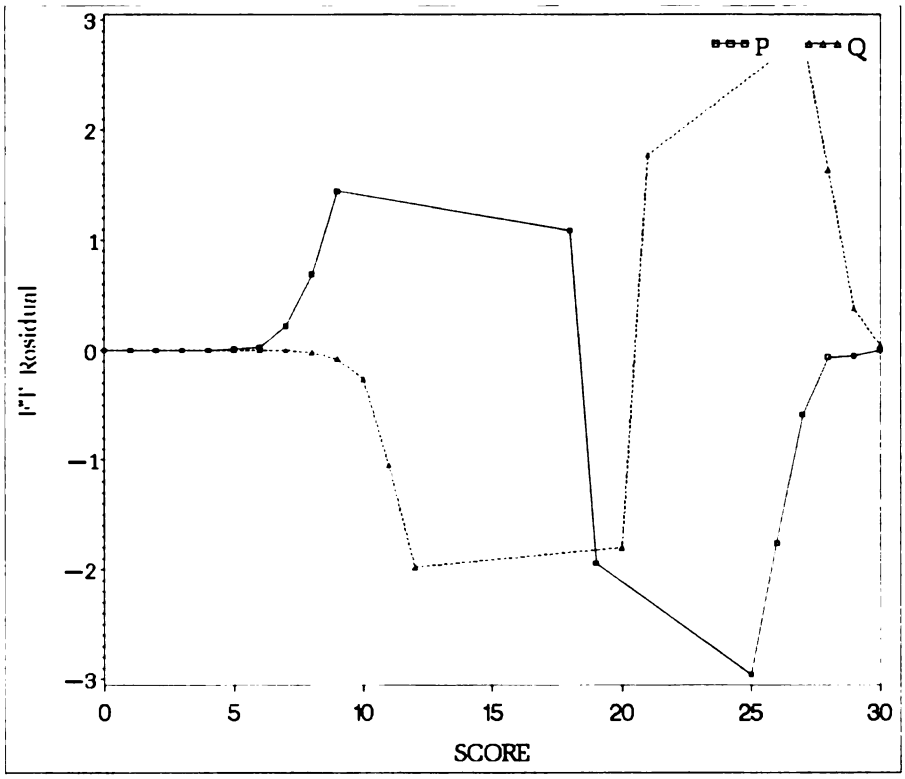

Figure A.25 Group Differences, 30 Missing Items, Anchor Test Score (A)

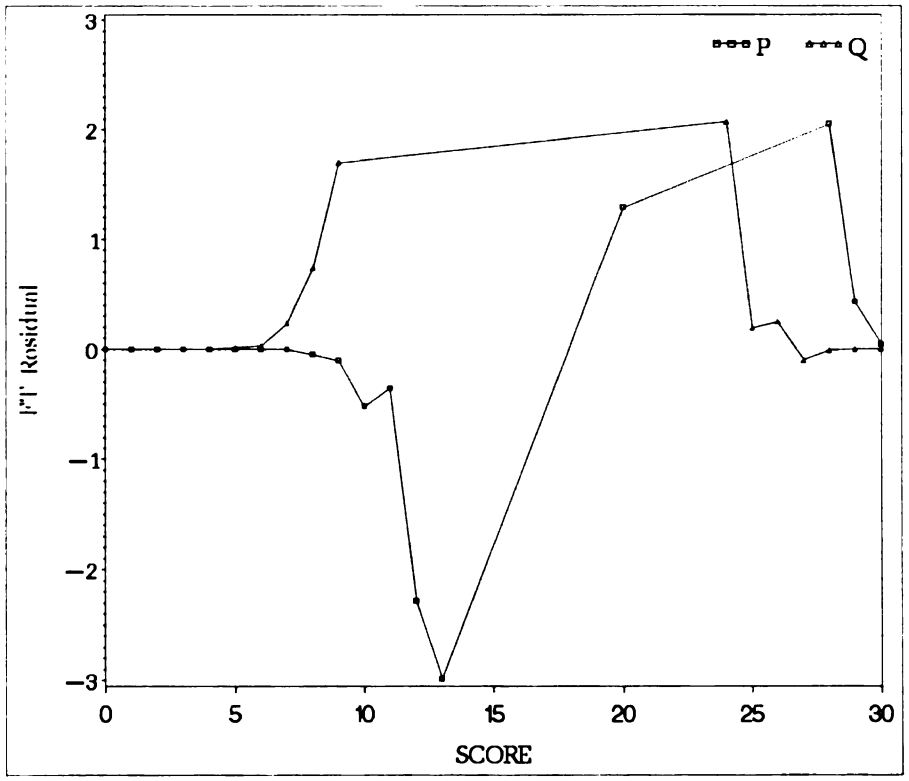

Figure A.26 Group Differences, 30 Missing Items, All Collateral Information (ALL)

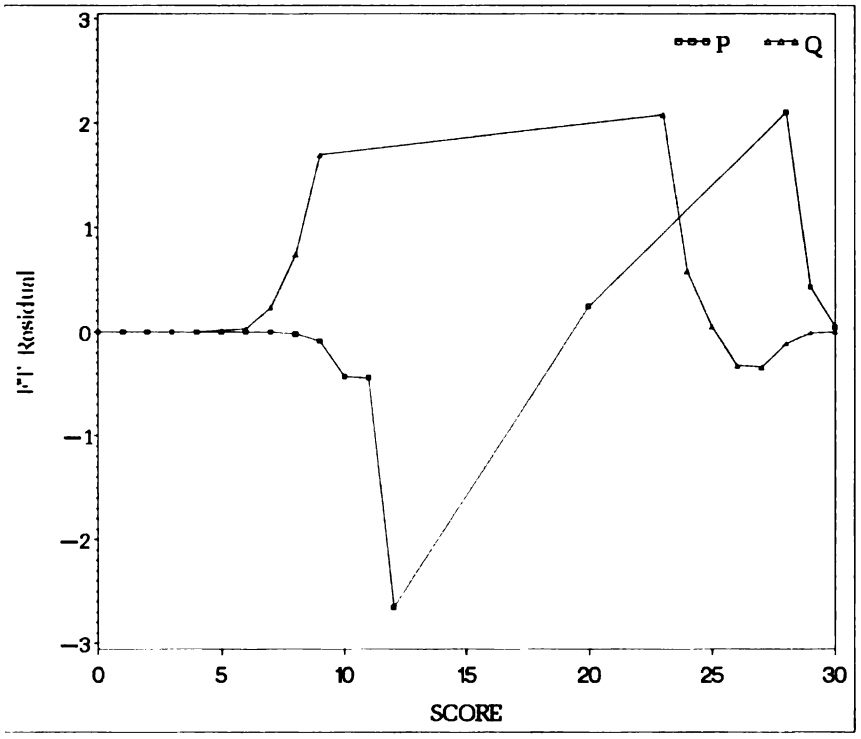

Figure A.27 Group Differences, 30 Missing Items, Anchor Test Score and Demographic Variables (A&D)

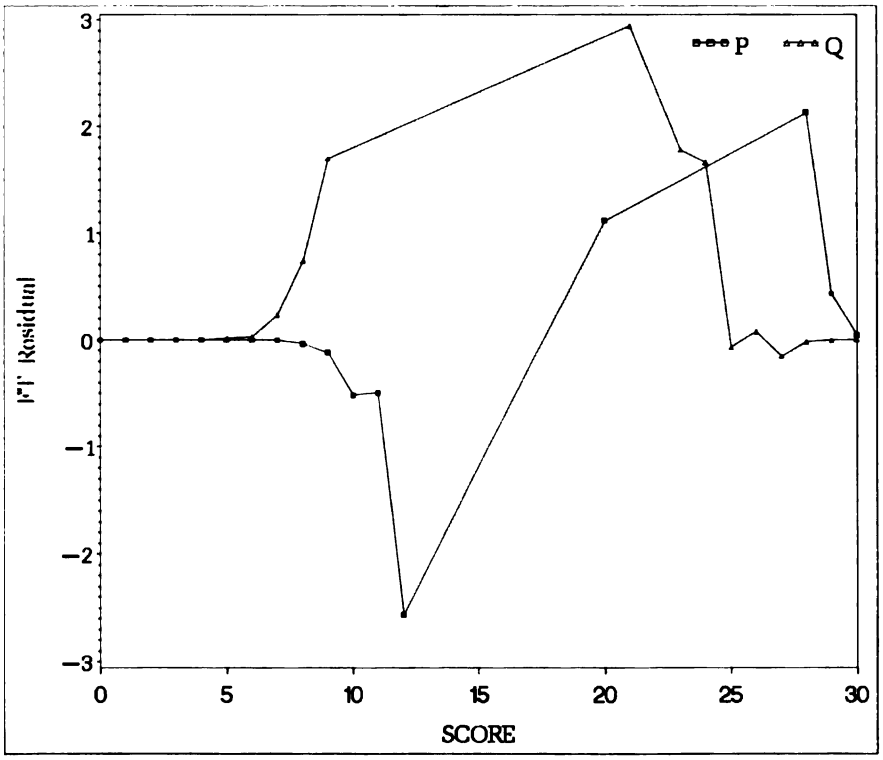

Figure A.28 Group Differences, 30 Missing Items, Subscores and Demographic Variables (S&D)

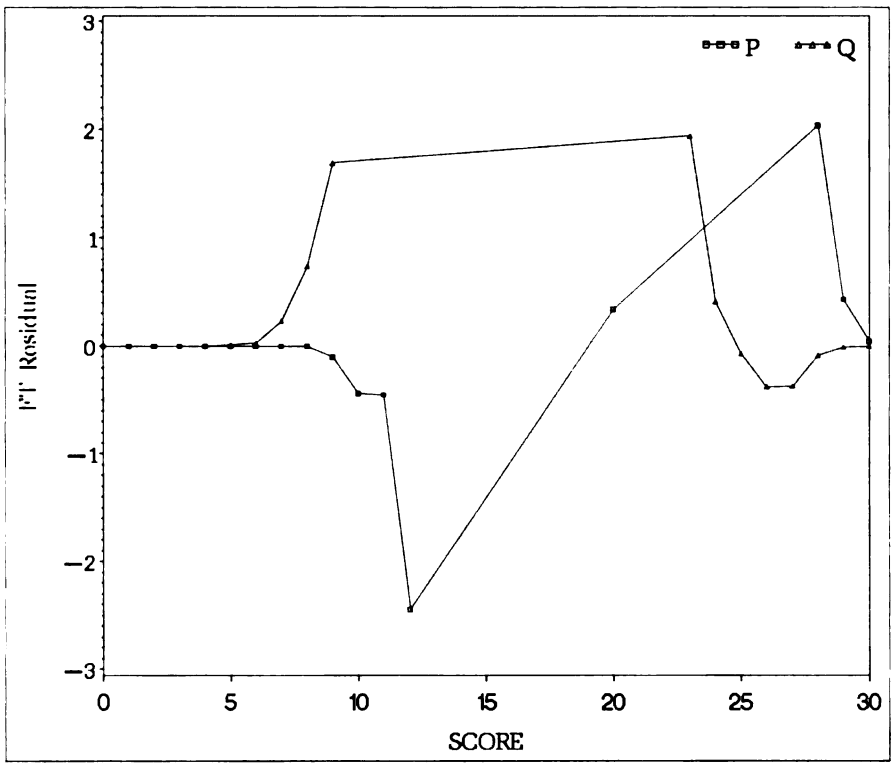

Figure A.29 Group Differences, 30 Missing Items, Demographic Variables (D)
S

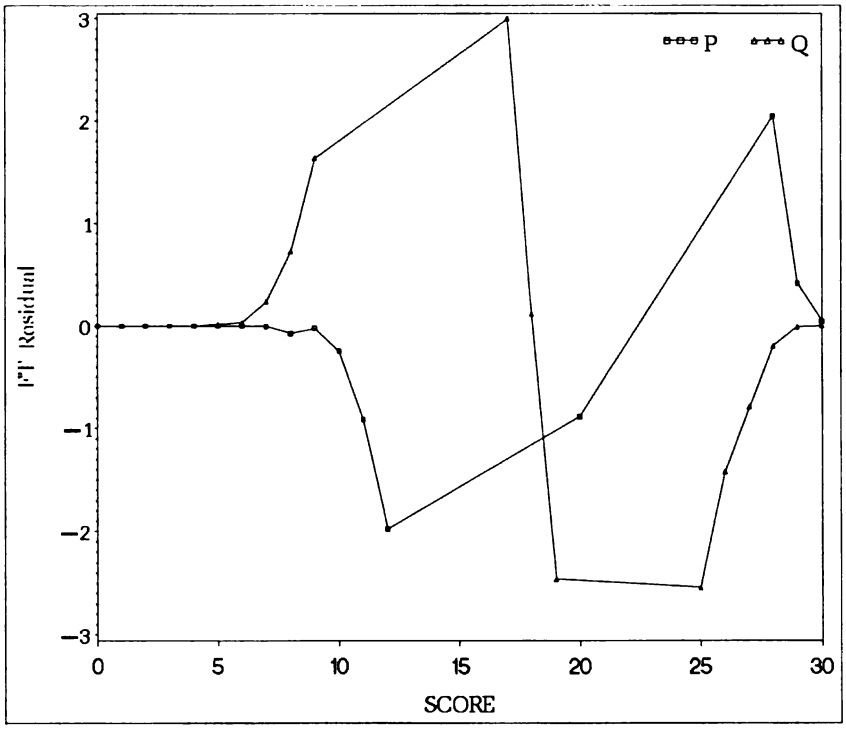

Figure A.30 Group Differences, 30 Missing Items, Subscores (S)

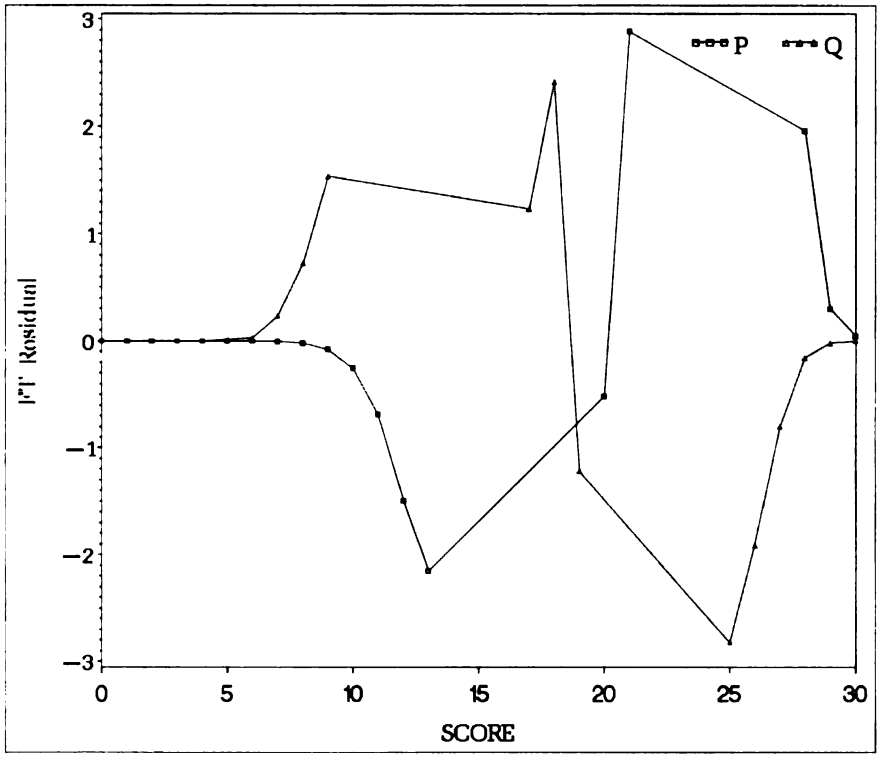

Figure A.31 Group Differences, 30 Missing Items, Anchor Test True Score (T)

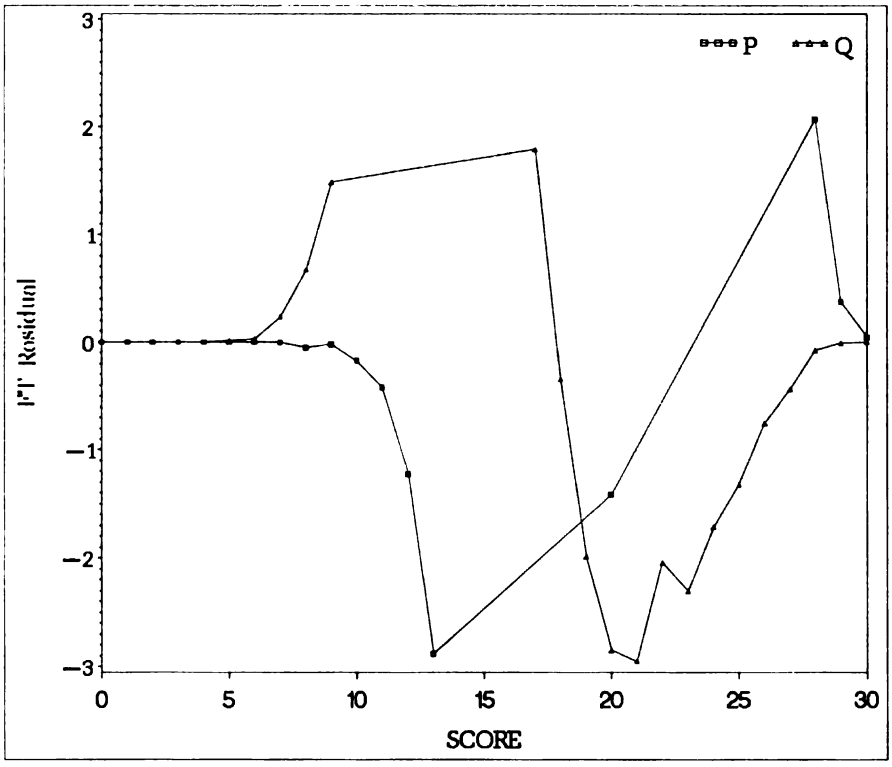

Figure A.32 Group Differences, 30 Missing Items, Subscores and Anchor Test Score (S&A)

Appendix B. FT Residuals for the Multiple Imputation Method

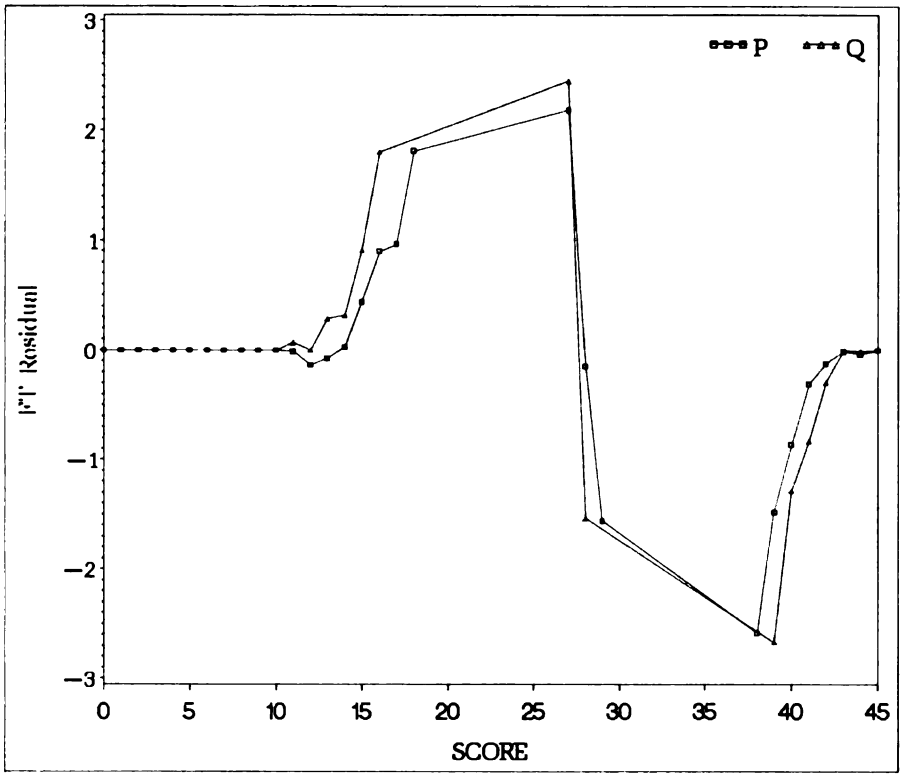

Figure B.1 No Group Differences, 45 Missing Items, Anchor Test Score (A)

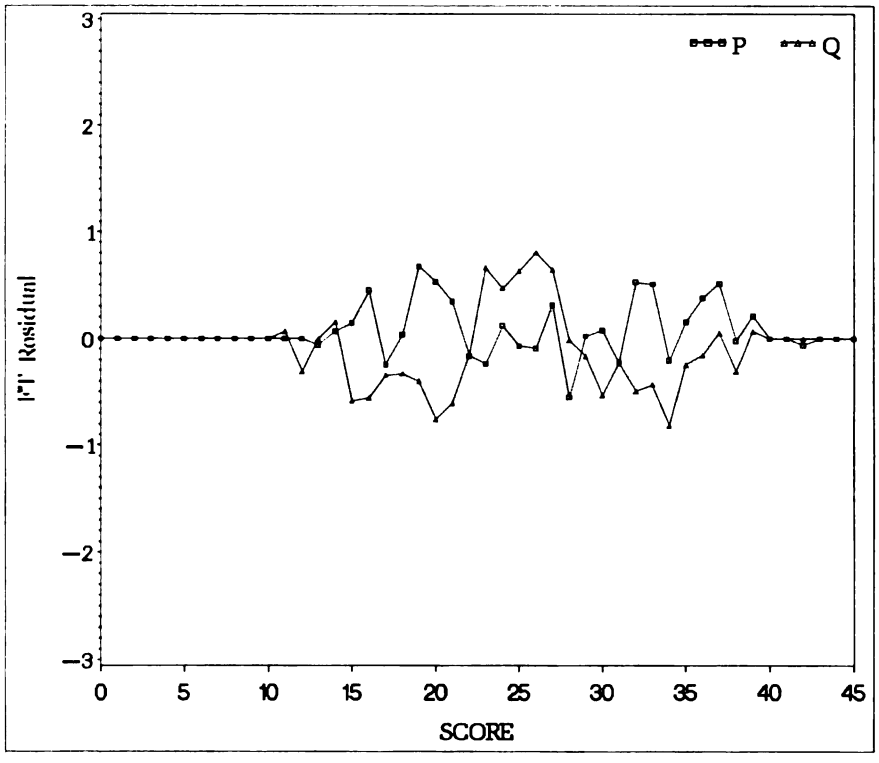

Figure B.2 No Group Differences, 45 Missing Items, All Collateral Information (ALL)

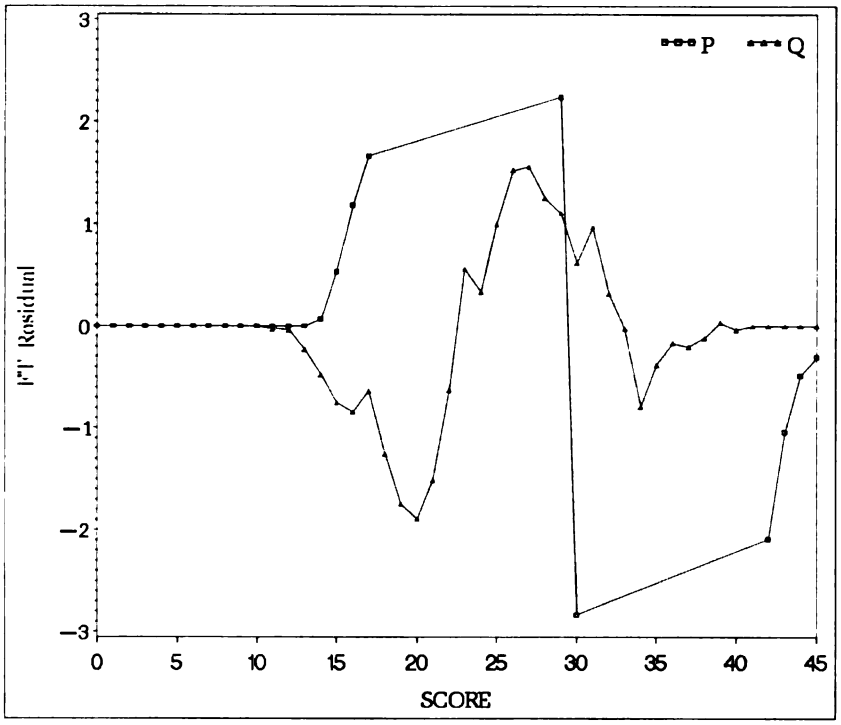

Figure B.3 No Group Differences, 45 Missing Items, Anchor Test Score and Demographic Variables (A&D)

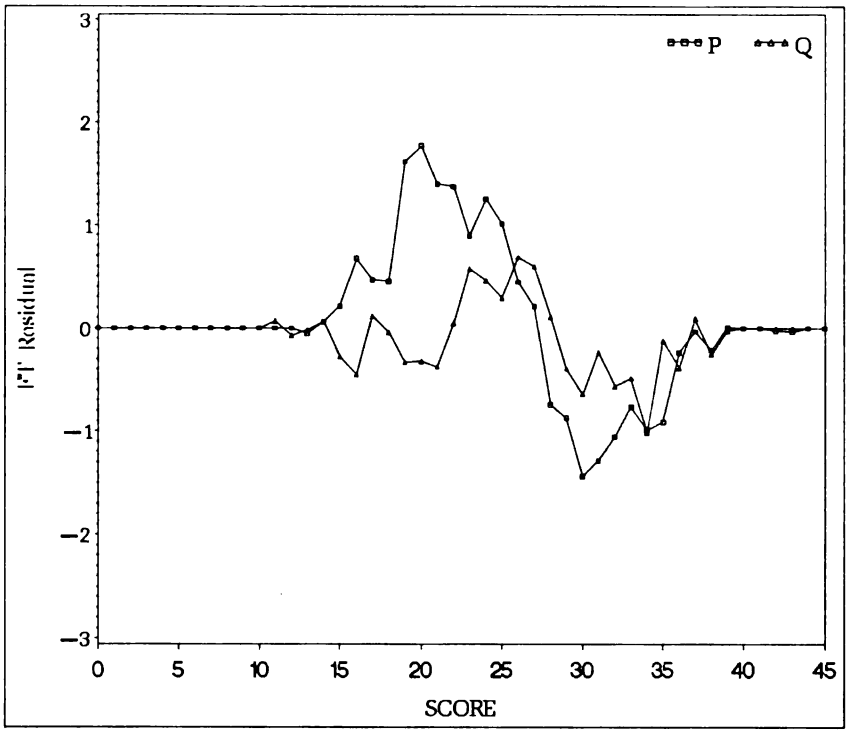

Figure B.4 No Group Differences, 45 Missing Items, Subscore and Demographic Variables (S&D)

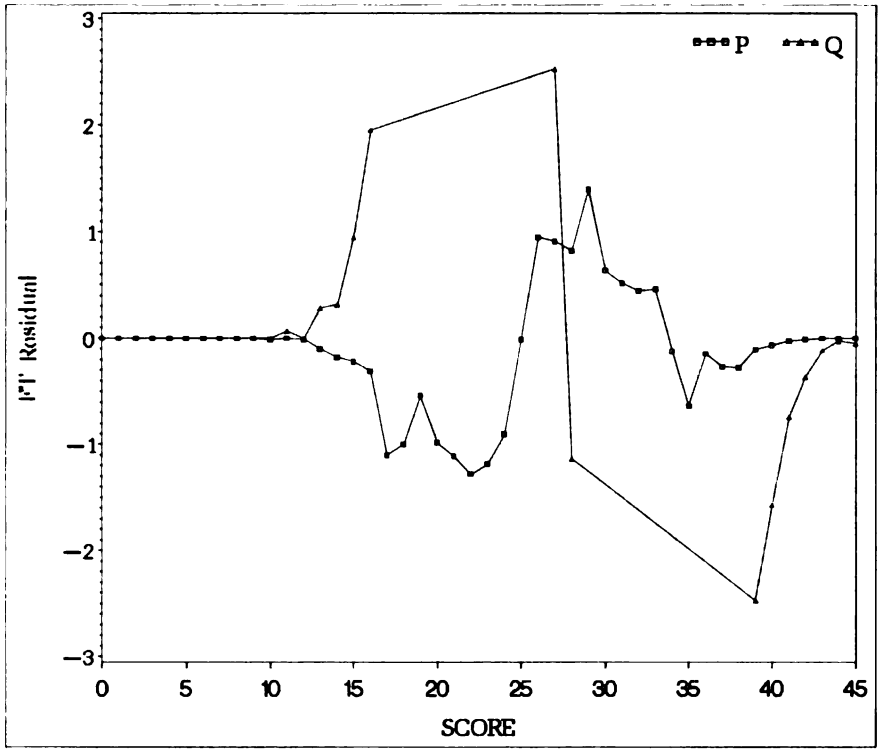

Figure B.5 No Group Differences, 45 Missing Items, Demographic Variables (D)

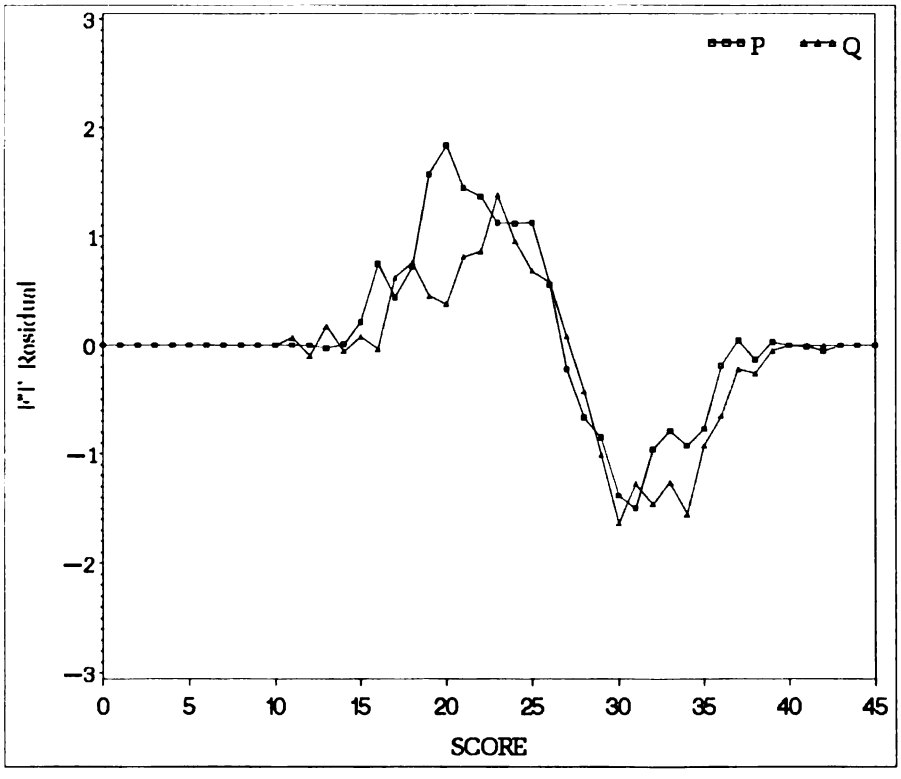

Figure B.6 No Group Differences, 45 Missing Items, Subscores (S)

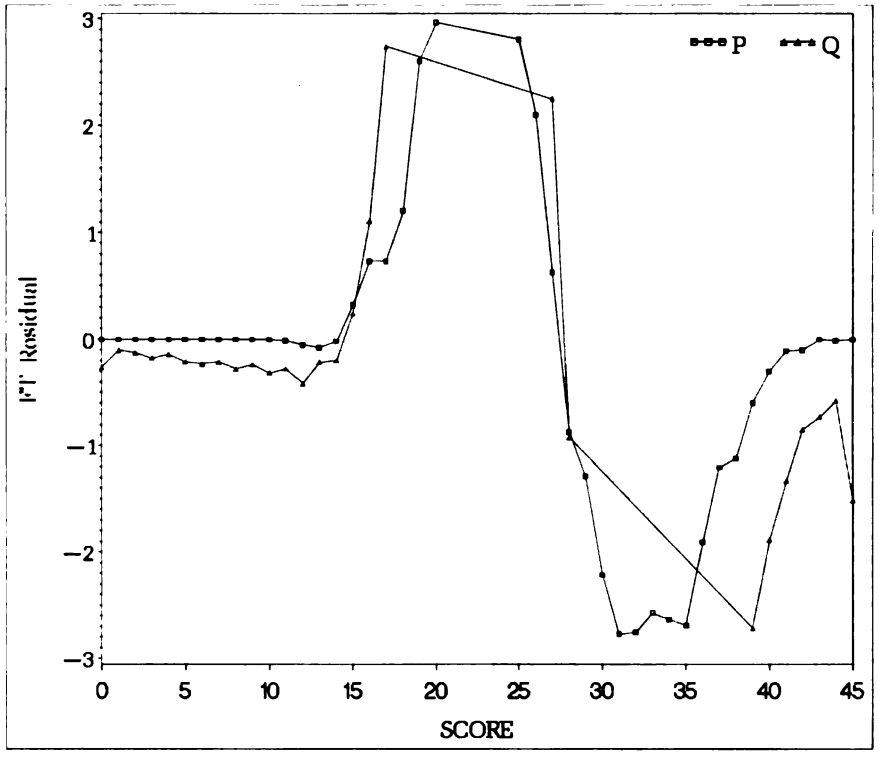

Figure B.7 No Group Differences, 45 Missing Items, Anchor Test True Score (T)

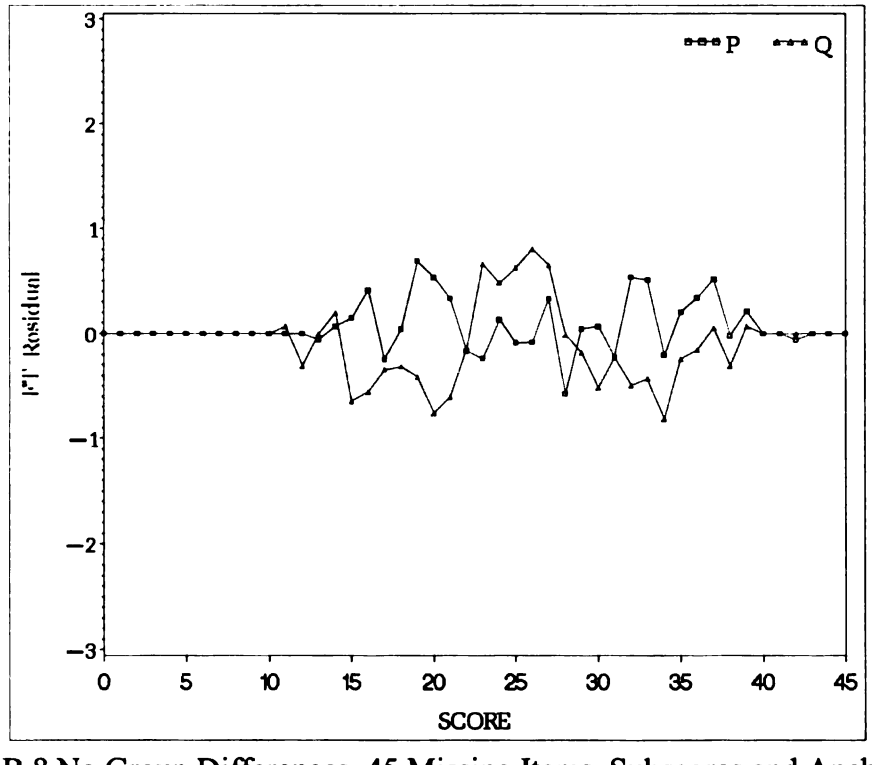

Figure B.8 No Group Differences, 45 Missing Items, Subscores and Anchor Test Score (S&A)

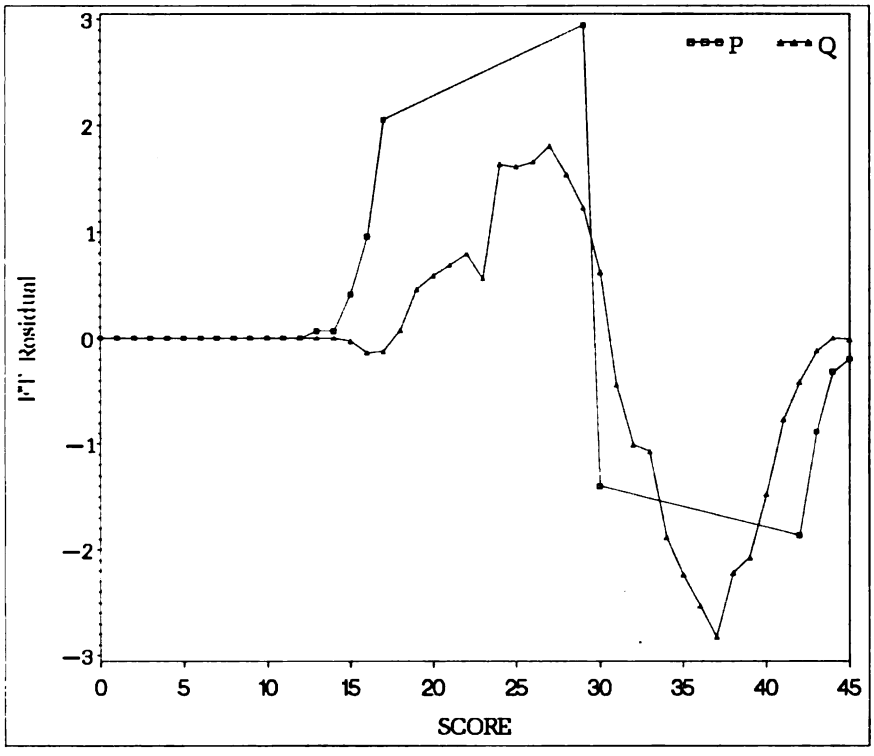

Figure B.9 Group Differences, 45 Missing Items, Anchor Test Score (A)

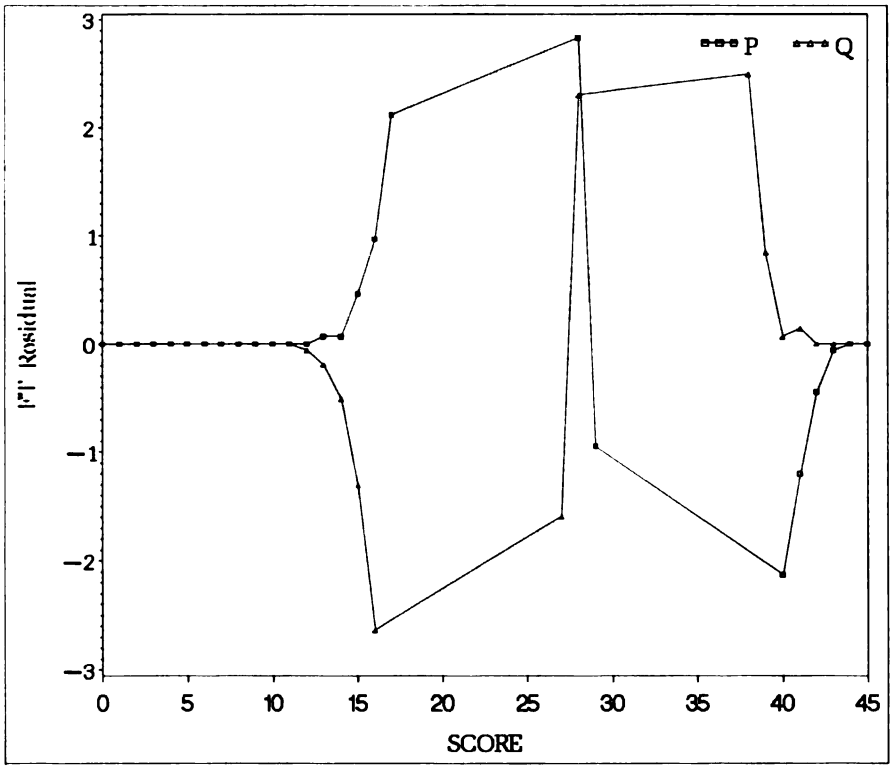

Figure B.10 Group Differences, 45 Missing Items, All Collateral Information (ALL)

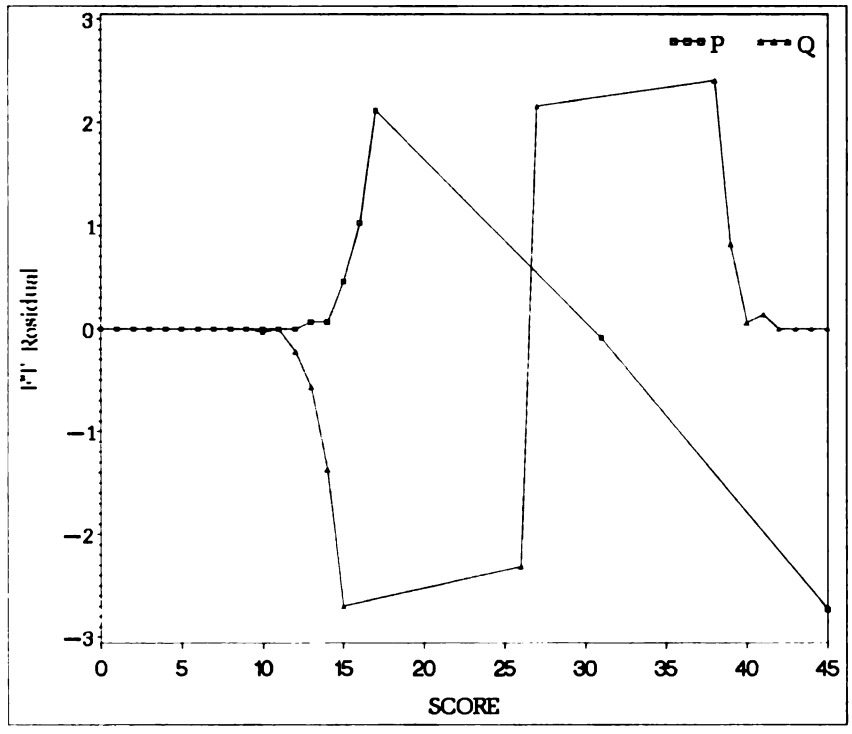

Figure B.11 Group Differences, 45 Missing Items, Anchor Test Score & Demographic variables (A&D)

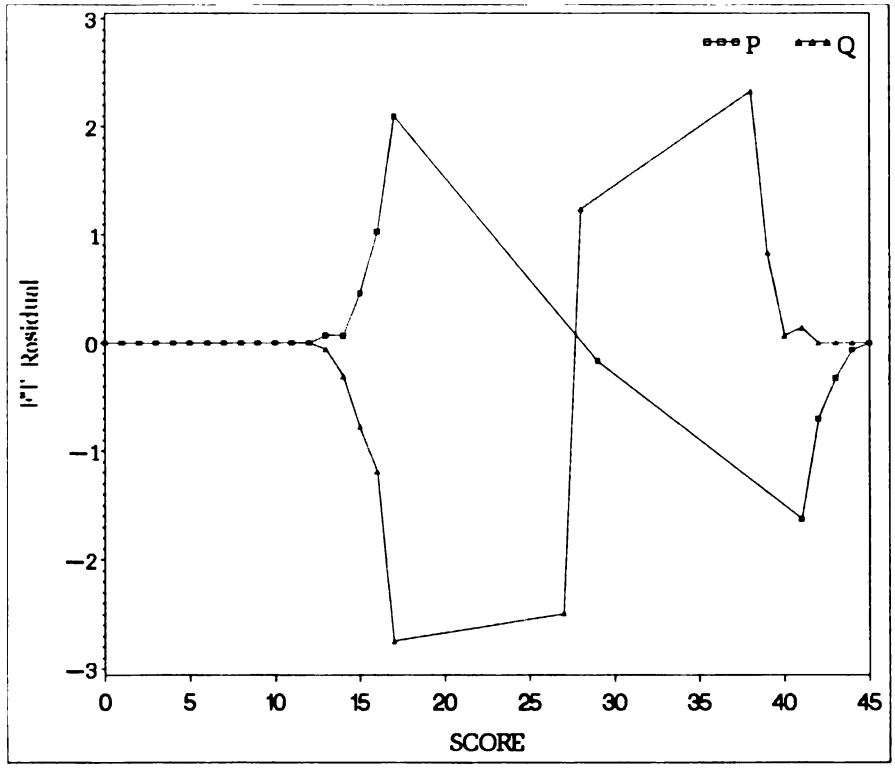

Figure B.12 Group Differences, 45 Missing Items, Subscores & Demographic variables (S&D)

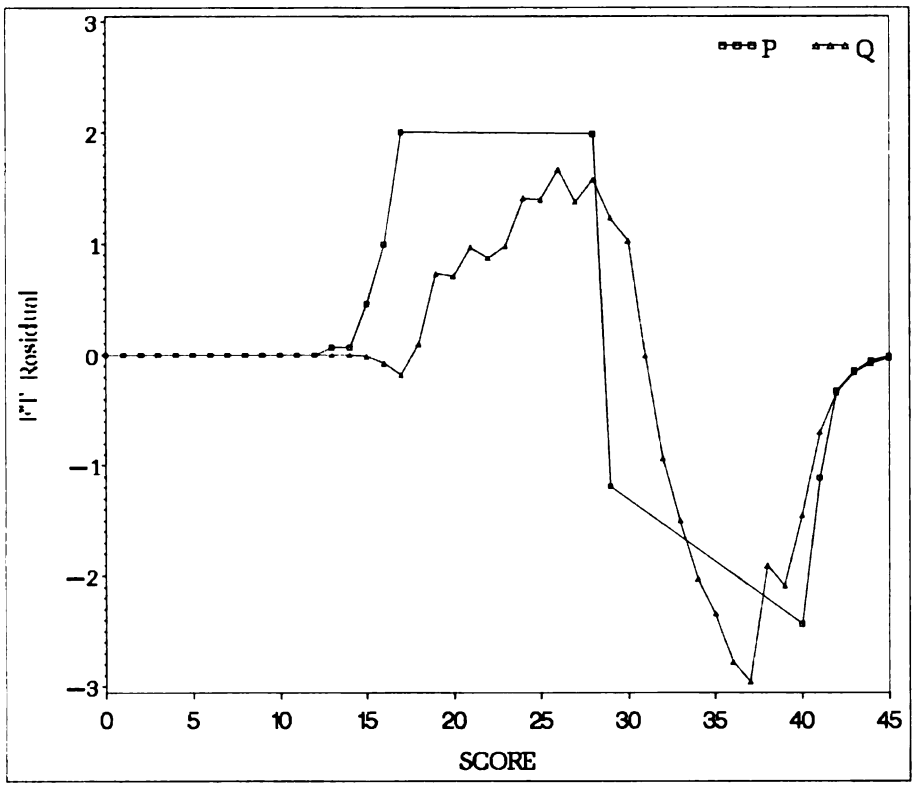

Figure B.13 Group Differences, 45 Missing Items, Demographic variables (D)

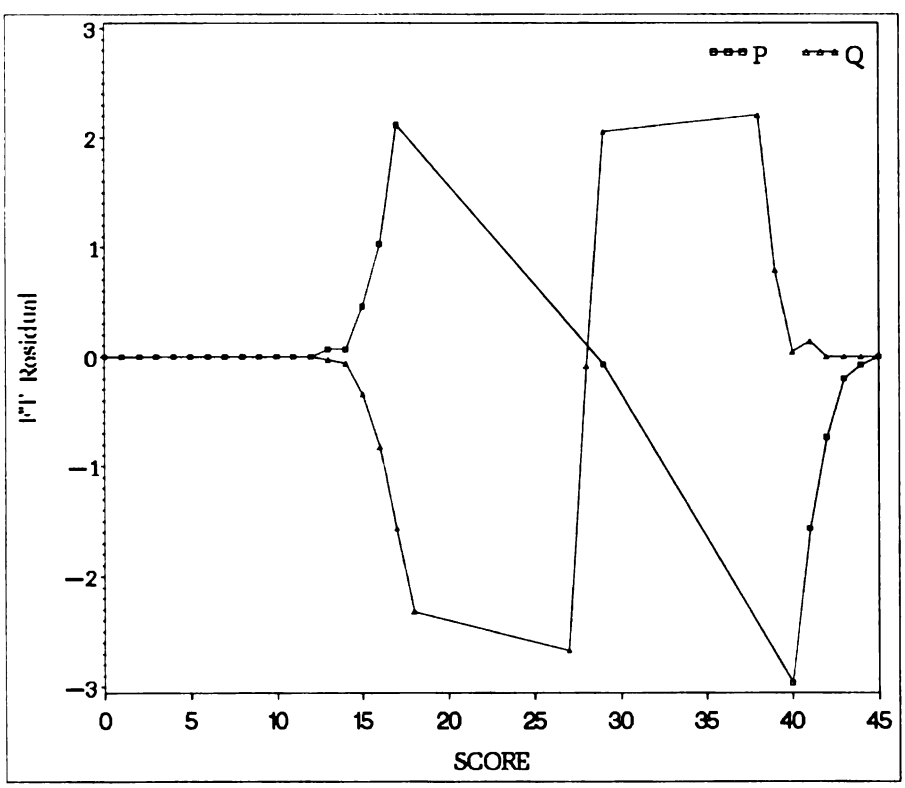

Figure B.14 Group Differences, 45 Missing Items, Subscores (S)

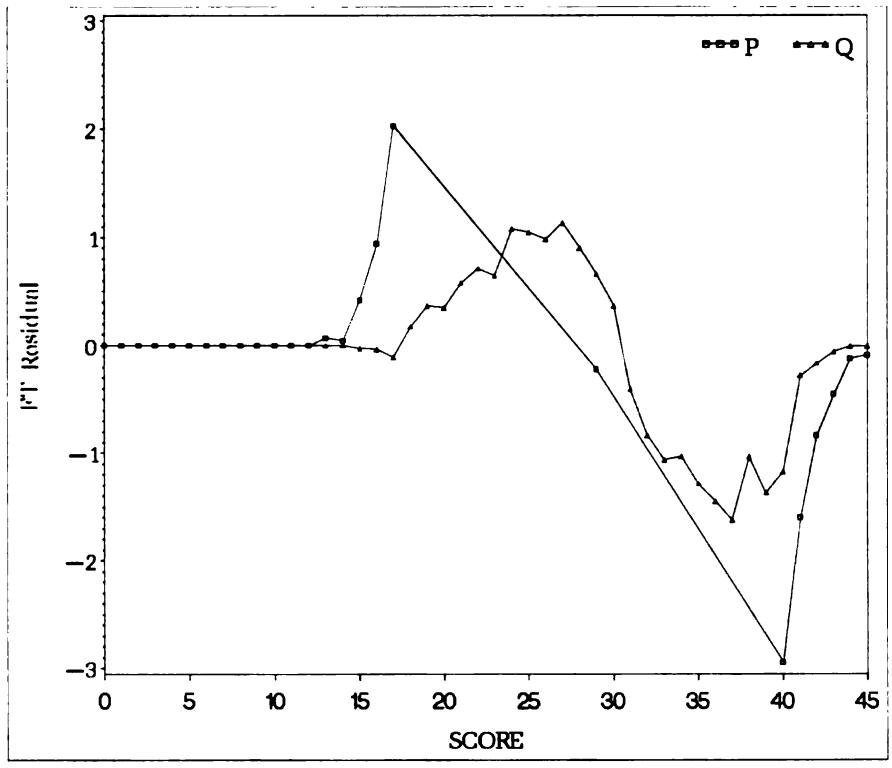

Figure B.15 Group Differences, 45 Missing Items, Anchor Test True Score (T)

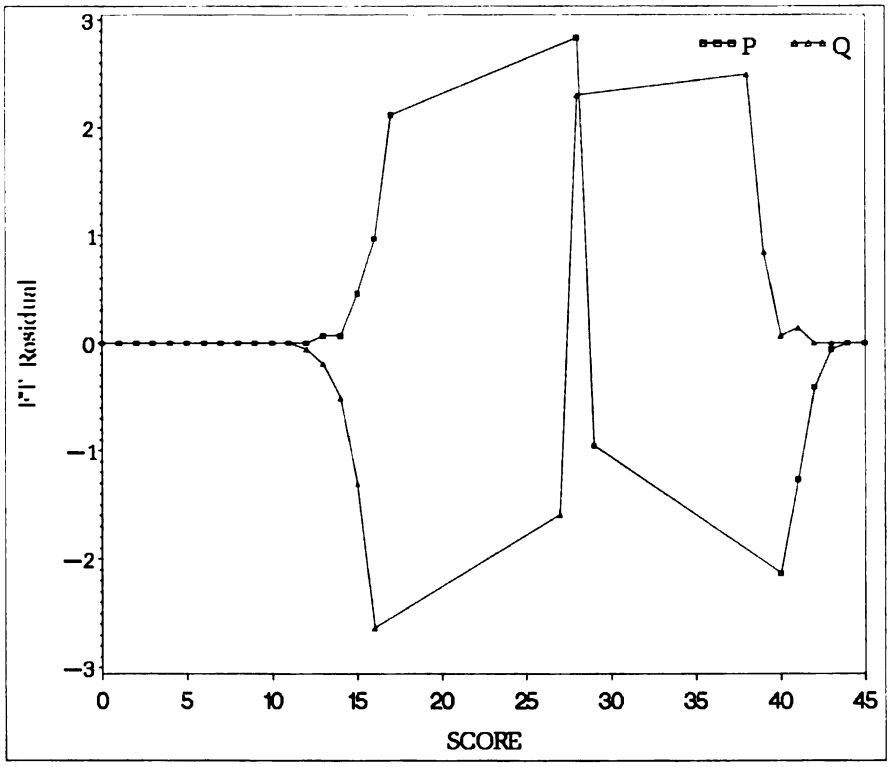

Figure B.16 Group Differences, 45 Missing Items, Subscores and Anchor Test Score (S&A)

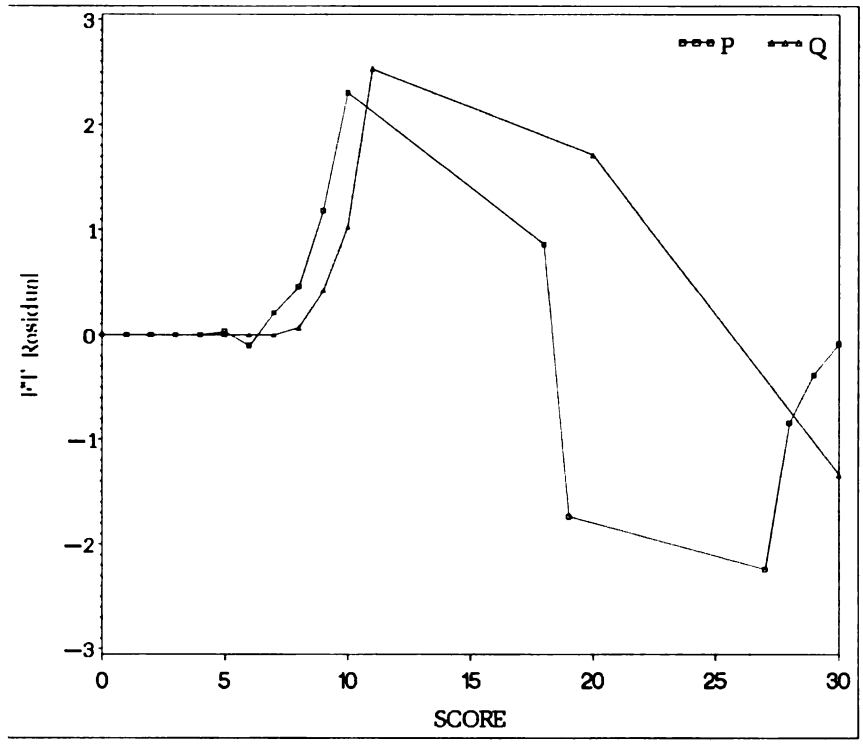

Figure B.17 No Group Differences, 30 Missing Items, Anchor Test Score (A)

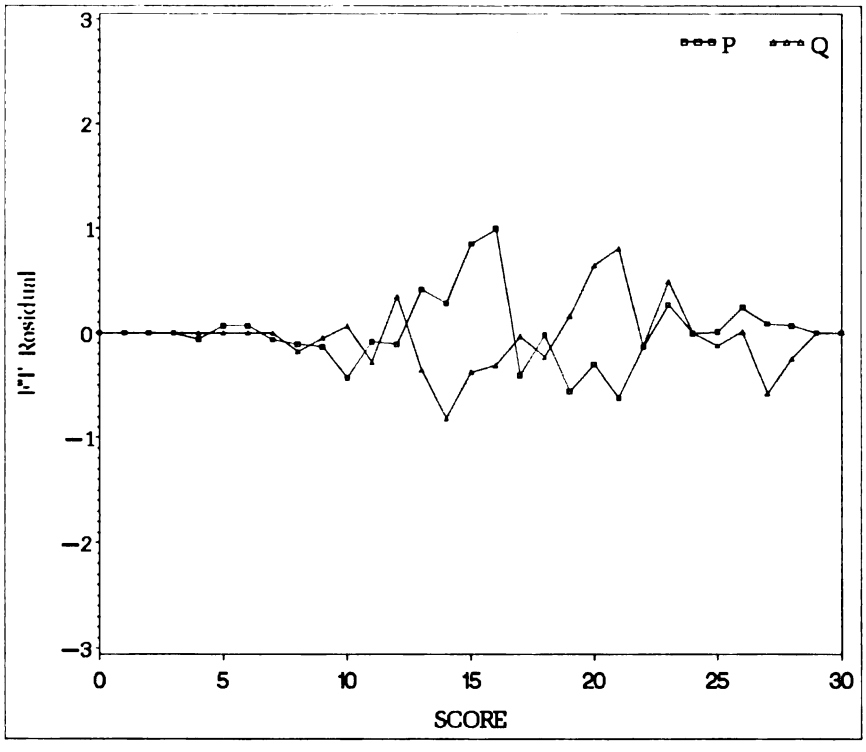

Figure B.18 No Group Differences, 30 Missing Items, All Collateral Information (ALL)

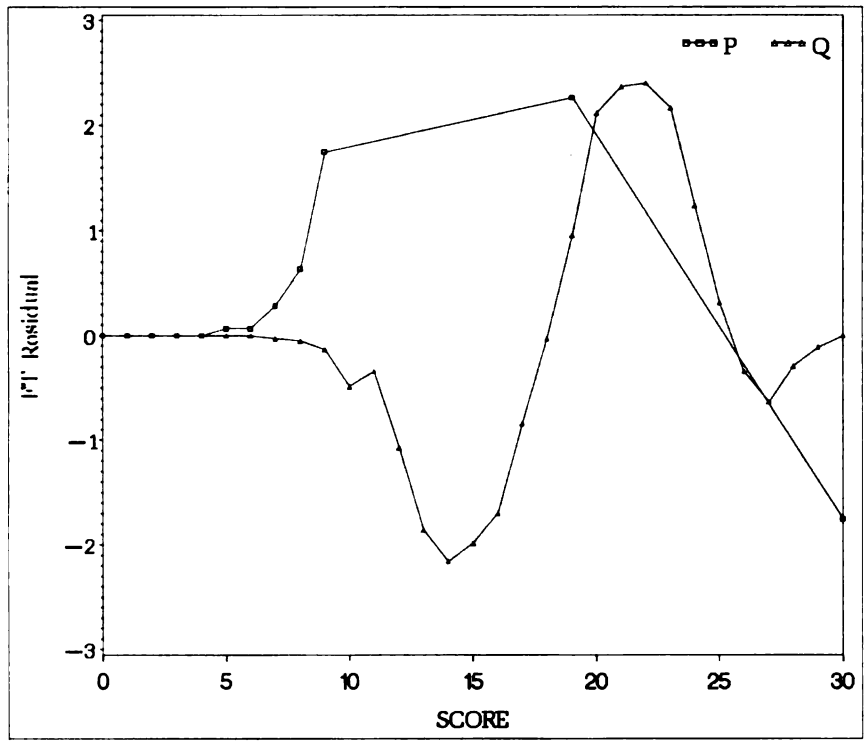

Figure B.19 No Group Differences, 30 Missing Items, Anchor Test Score and Demographic Variables (A&D)

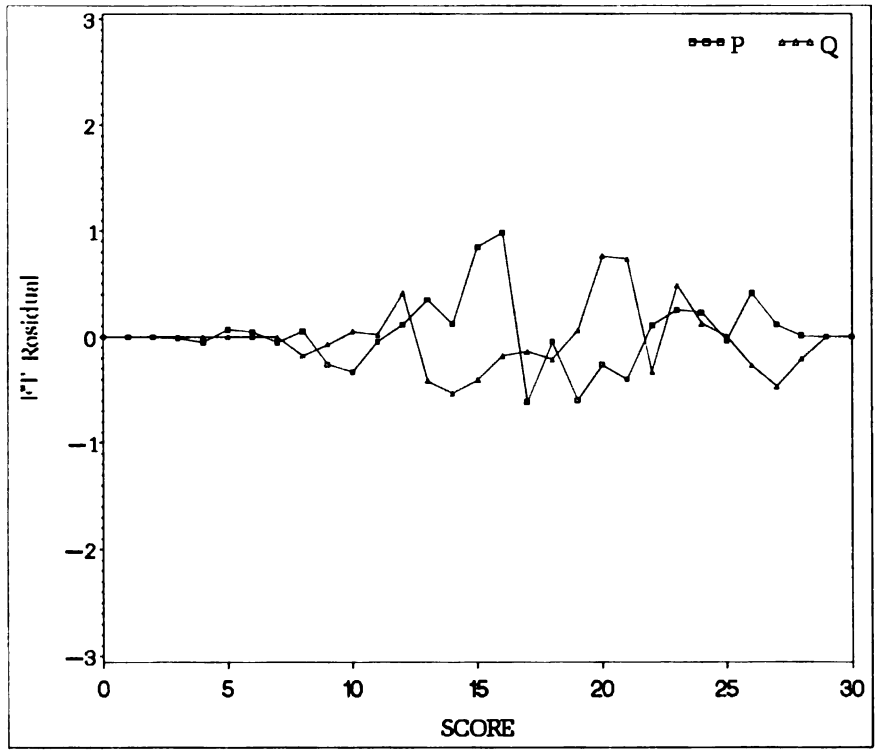

Figure B.20 No Group Differences, 30 Missing Items, Subscores and Demographic Variables (S&D)

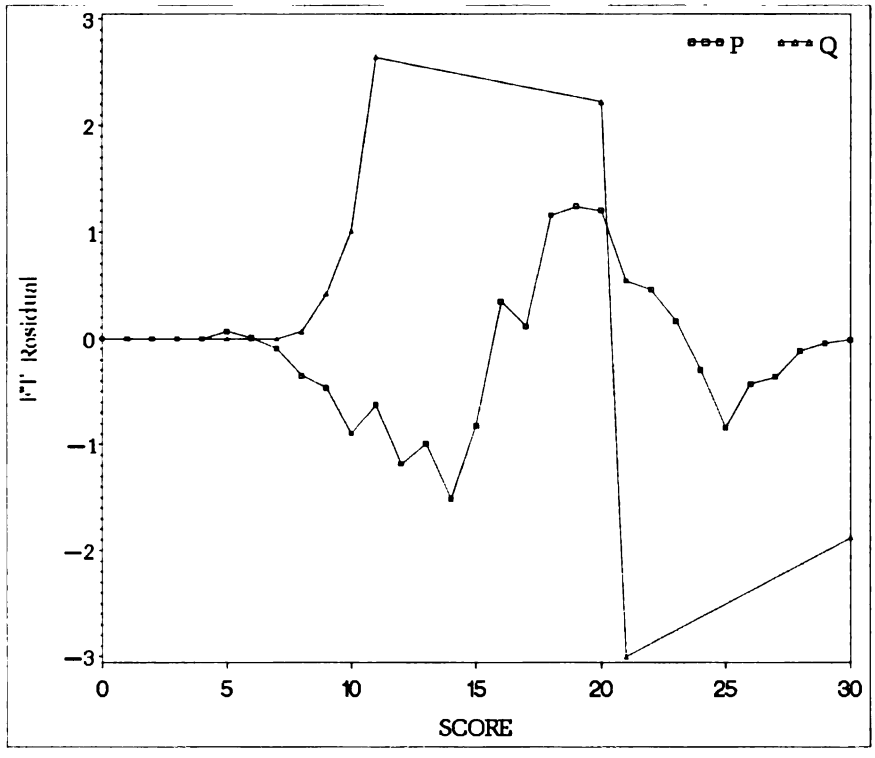

Figure B.21 No Group Differences, 30 Missing Items, Demographic Variables (D)

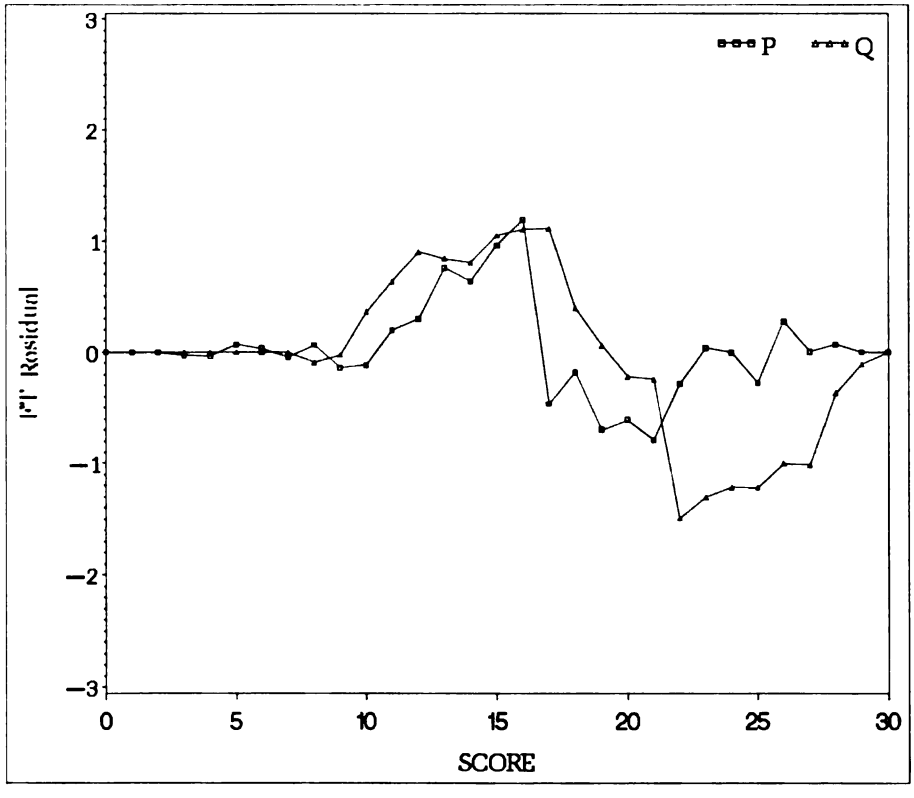

Figure B.22 No Group Differences, 30 Missing Items, Subscores (S)

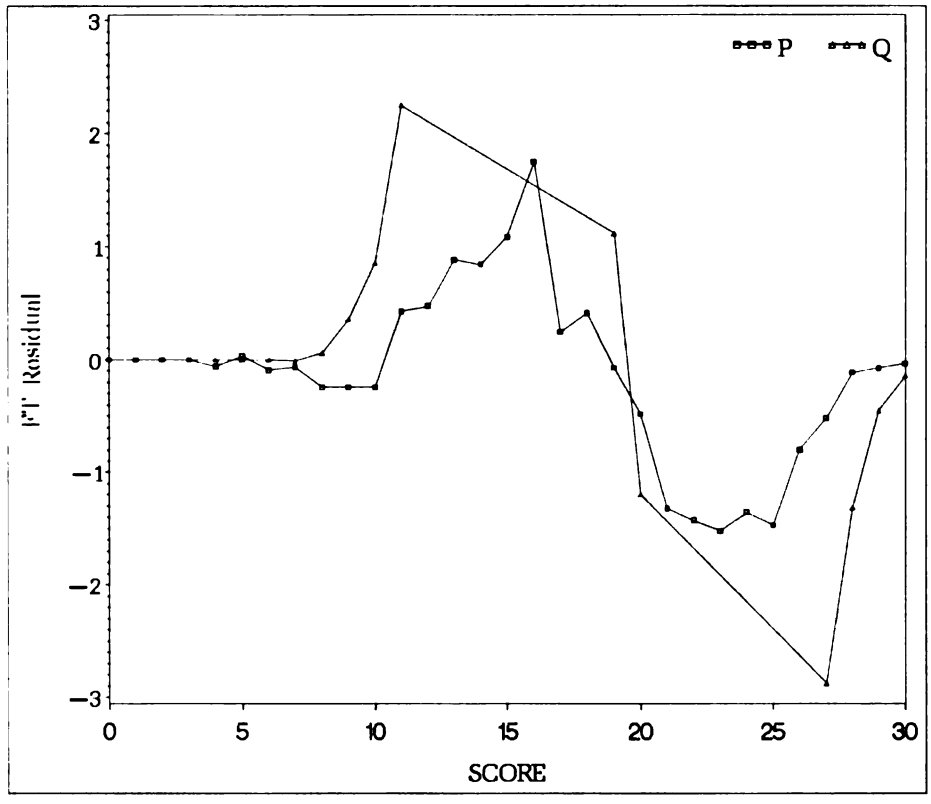

Figure B.23 No Group Differences, 30 Missing Items, Anchor Test True Score (T)

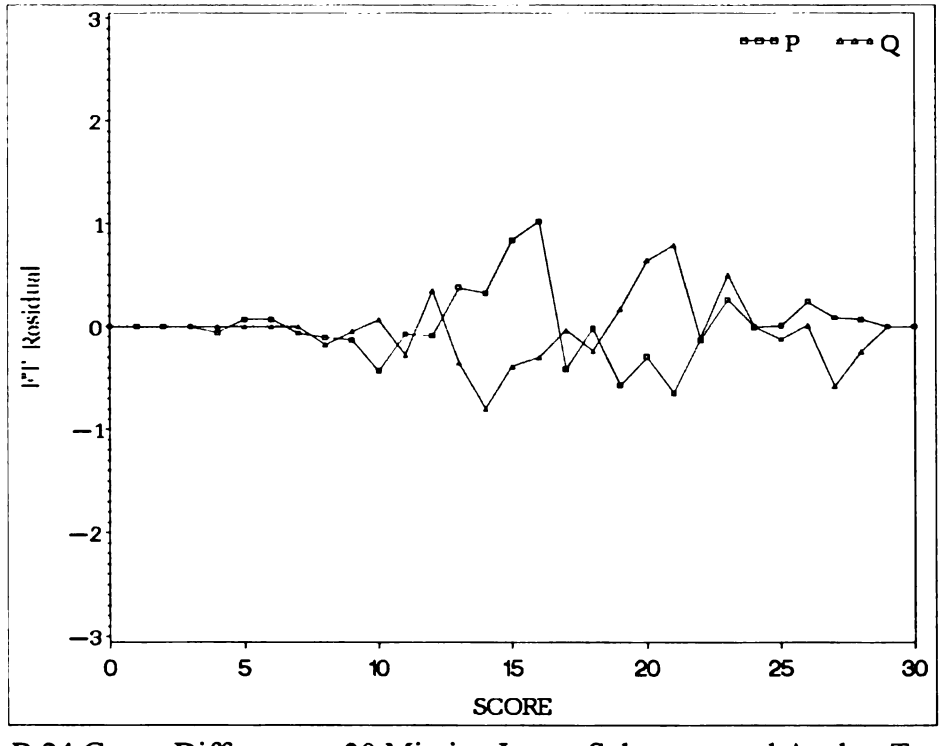

Figure B.24 Group Differences, 30 Missing Items, Subscores and Anchor Test Score (S&A)

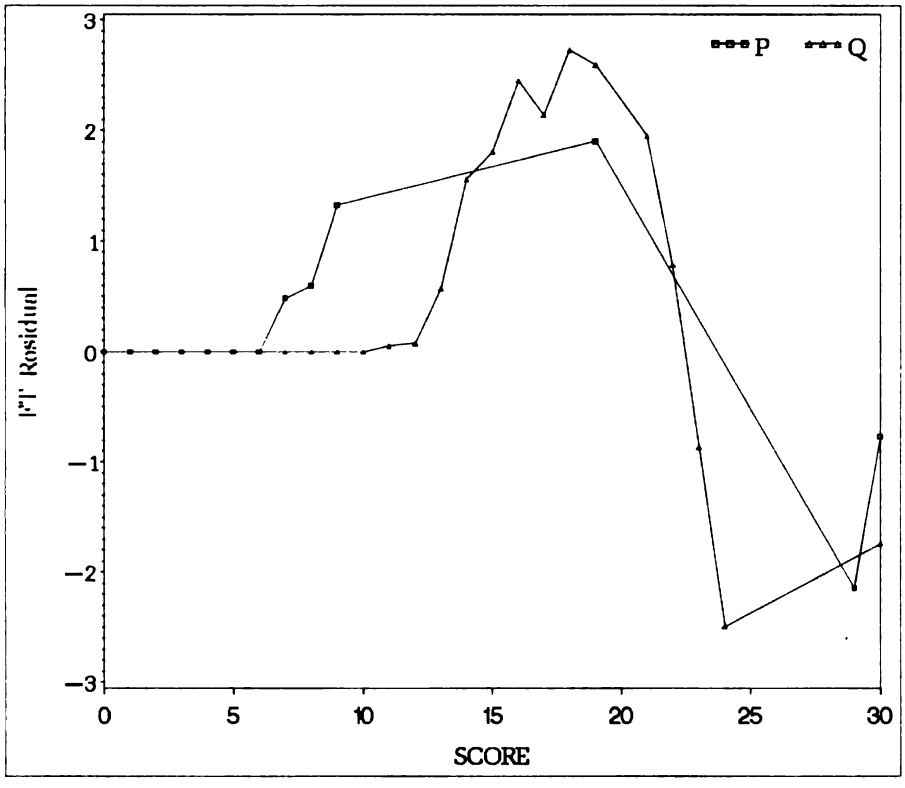

Figure B.25 Group Differences, 30 Missing Items, Anchor Test Score (A)

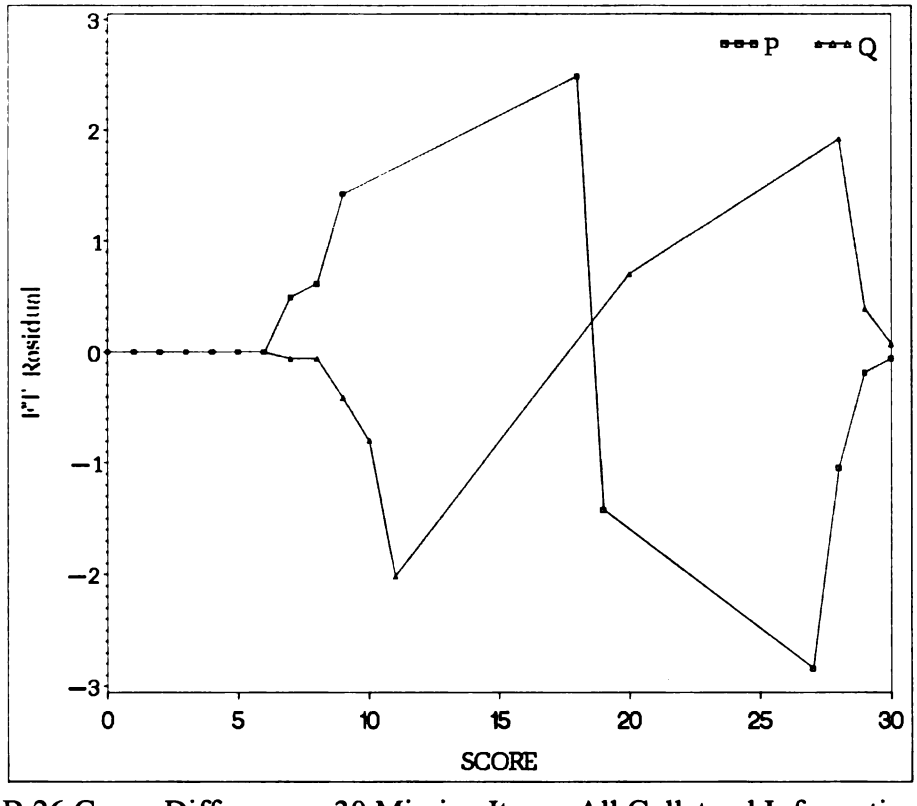

Figure B.26 Group Differences, 30 Missing Items, All Collateral Information (ALL)

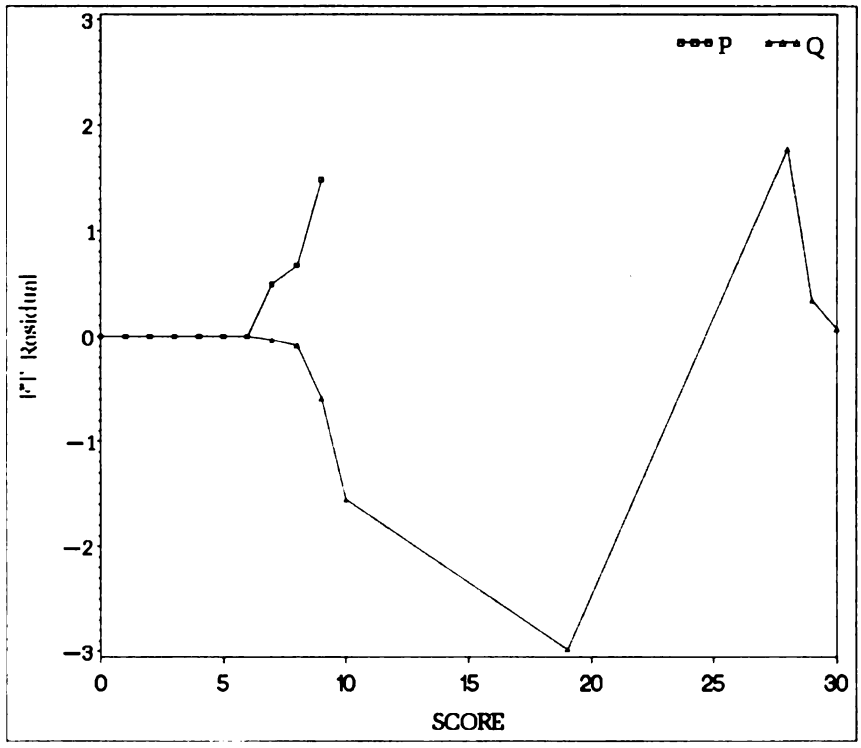

Figure B.27 Group Differences, 30 Missing Items, Anchor Test Score and Demographic Variables (A&D)

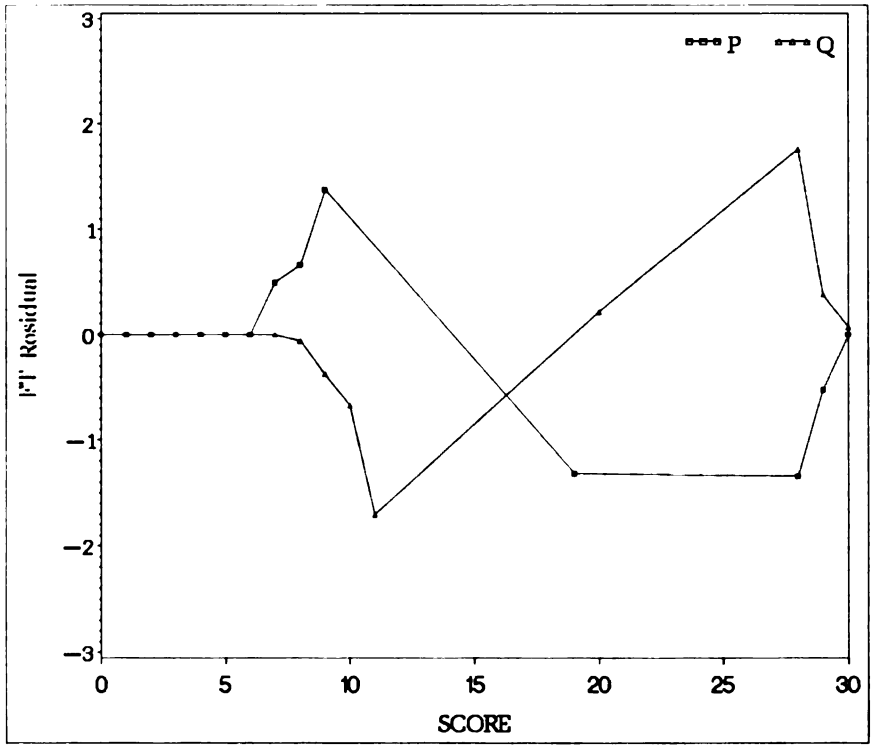

Figure B.28 Group Differences, 30 Missing Items, Subscores and Demographic Variables (S&D)

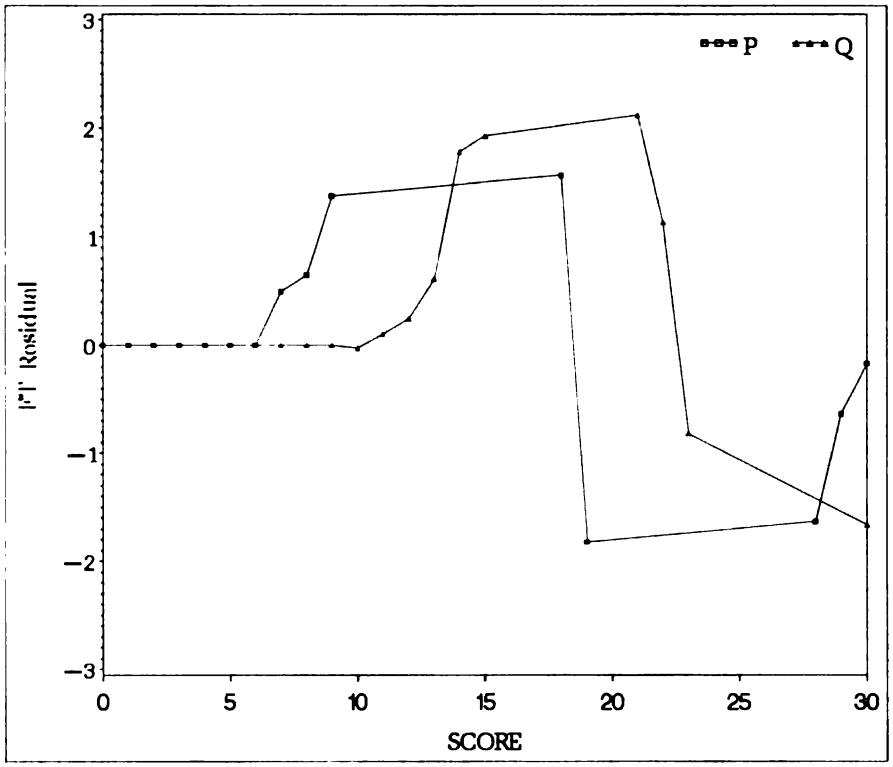

Figure B.29 Group Differences, 30 Missing Items, Demographic Variables (D)

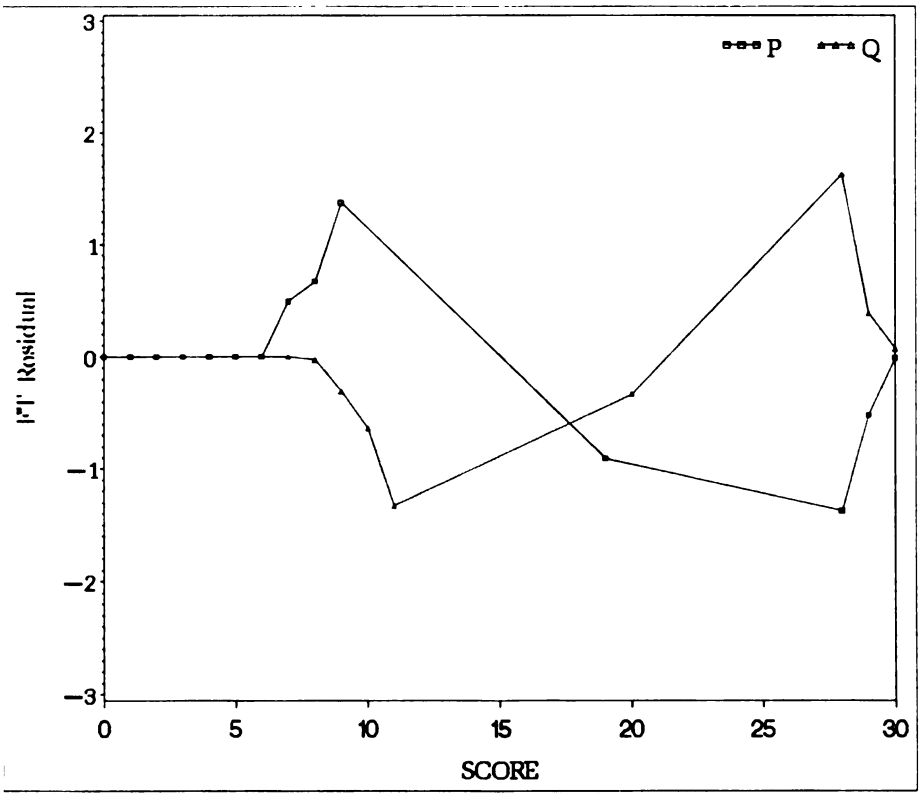

Figure B.30 Group Differences, 30 Missing Items, Subscores (S)

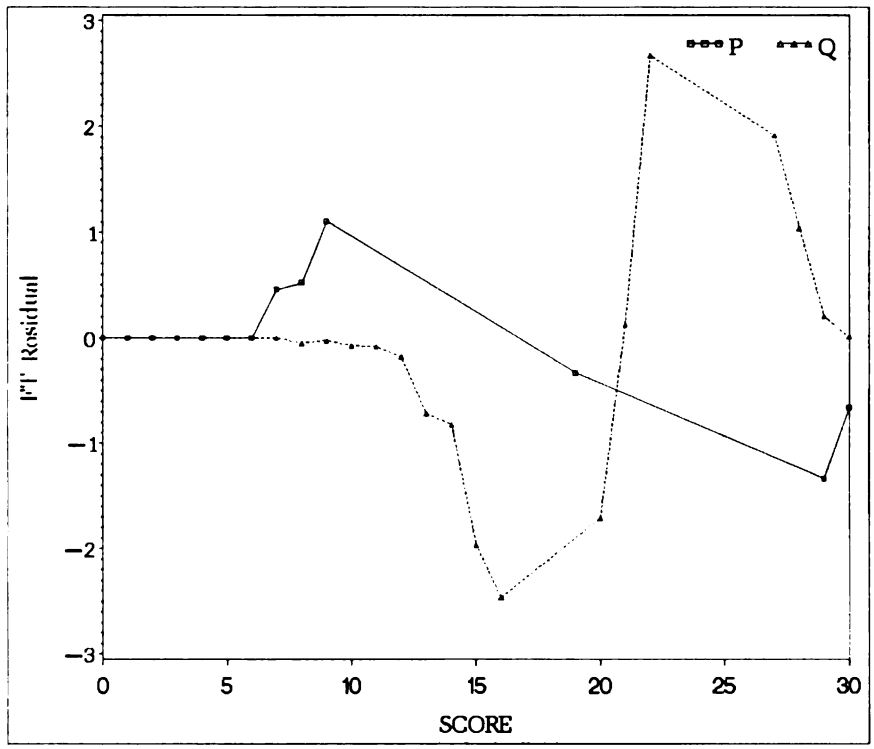

Figure B.31 Group Differences, 30 Missing Items, Anchor Test True Score (T)

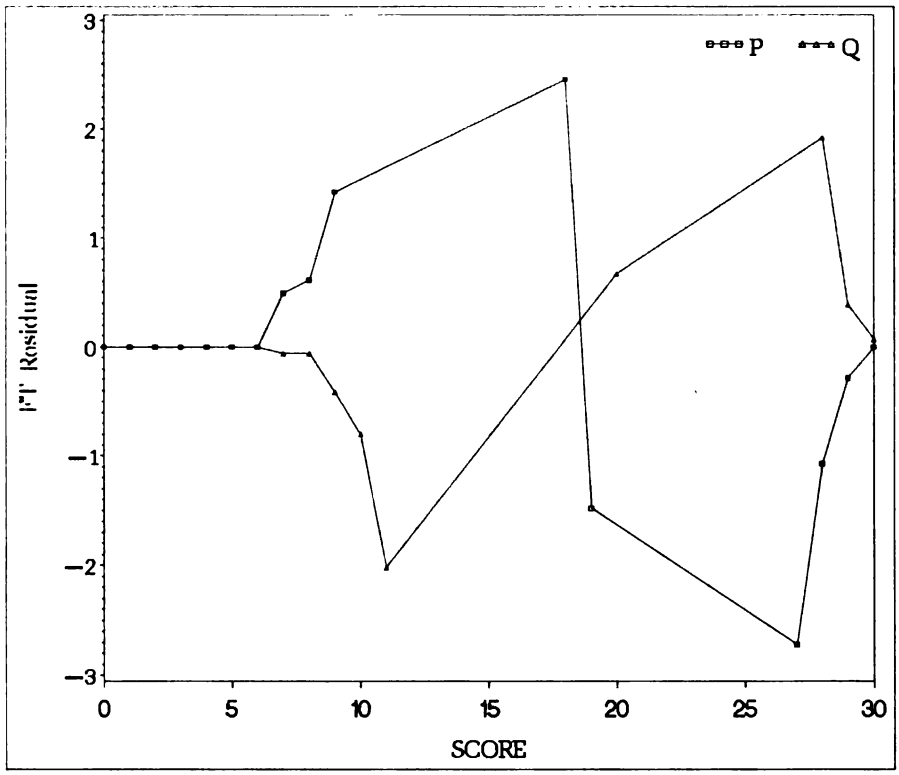

Figure B.32 Group Differences, 30 Missing Items, Subscores and Anchor Test Score (S&A)

APPENDIX C: WinBUGS Code for the Test Form 1

```
Model
{
#Read in Item Responses
for (j in 1:N){
for (k in 1:T){
x[j,k] <- response[j,k]
}}
# Specify five-dimensional Two-Parameter Logistic Model
for (j in 1:N) {
for (k in 1:T) {
p[j,k] <-
exp(a1[k]*theta[j,1]+a2[k]*theta[j,2]+a3[k]*theta[j,3]+a4[k]*theta[j,4]+a5[k]*theta[j,5]+
d[k])/
(1+exp(a1[k]*theta[j,1]+a2[k]*theta[j,2]+a3[k]*theta[j,3]+a4[k]*theta[j,4]+a5[k]*theta[j,
5]+d[k]))
x[j,k] ~ dbern(p[j,k])
}
#Specify Prior for Thetas
theta[j,1:5] ~ dmnorm(mu[1:5],tau[1:5,1:5])
}
# Specify Priors for Discrimination and Difficulty Parameters
a1[1] <-0
a2[1] ~ dnorm(0,.5)
a3[1] ~ dnorm(0,.5)
a4[1] ~ dnorm(0,.5)
a5[1] ~ dnorm(0,.5)
d[1] ~ dnorm(0,.5)
for (k in 2:62) {
a1[k] ~ dnorm(0,.5);
a2[k] ~ dnorm(0,.5);
a3[k] ~ dnorm(0,.5);
a4[k] ~ dnorm(0,.5);
a5[k] ~dnorm(0,.5);
d[k] ~ dnorm(0,.5);
}}
list(N=1361,T=62, mu=c(0,0,0,0,0),
tau = structure(.Data = c(1,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,1),.Dim = c(5,5)),
response=structure(.Data=c(
0,1,1,1,0,0,0,0,0,0,0,0,0,1,0,1,0,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0,1,0,1,0,1,1,1,0,1,0,0,1,1,0,0,1,0,0,0,0,0,1,1,0,0,0,0,0,0,0,1,
1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,1,1,1,0,1,1,0,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,
1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,0,1,0,1,0,0,1,0,0,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,0,0,0,0,
1,0,0,0,0,0,0,0,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,1,1,1,0,0,1,0,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,1,1,1,0,1,0,0,
1,0,1,1,1,0,1,1,1,1,1,1,1,0,0,1,1,0,1,0,1,1,0,1,1,1,0,1,0,1,1,1,1,1,0,0,1,0,1,1,0,1,0,0,1,0,1,1,0,0,0,1,0,0,0,0,1,1,1,1,0,1,1,1,
0,1,0,0,1,0,1,1,1,1,1,0,1,0,0,0,1,1,0,1,0,1,1,1,0,1,1,0,1,1,1,1,1,0,0,0,0,0,0,0,1,1,1,0,0,0,0,1,1,1,0,0,1,0,0,0,0,0,0,1,1,1,1,0,
```

0,1,1,1,1,1,1,1,1,1,1,1,0,1,0,1,0,0,0,0,0,0,1,1,0,1,1,1,1,1,0,0,1,0,1,1,0,0,1,1,0,0,0,0,0,0,1,0,1,0,1,1,0,1,1,0,1,1,0,1,1,0,
0,0,1,1,1,0,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,0,1,0,1,1,1,1,1,0,1,0,1,1,0,1,1,1,1,1,1,0,0,0,0,0,0,0,0,1,1,0),
.Dim=c(1361,62))

REFERENCES

REFERENCES

Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley & Sons.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington DC.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov Chain Monte Carlo. *Applied psychological measurement, 27(6)*, 395-414.

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

Chen, H., Yan, D., Hemat, L., Han, N., & von Davier, A. A. (2008). KE & Loglin Software v. 3.0[Computer Software]. Educational Testing Services, New Jersey.

Clauser, B. E., Margolis, M. J., & Case, S. M. (2006). Testing for licensure and certification in the professions. In R. L. Brennan (Ed.), *Educational Measurement*(4^{th}). New York: American Council on Education, Praeger.

von Davier, A. A. (2003). *Notes linear equating methods for the Non-Equivalent Groups designs* (ETS RR-03-24). Princeton, NJ: Educational Testing Services.

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

D'Agostino, R. B., & Rubin, D. R. (2006). In D. B. Rubin (Ed.). *Matched sampling for causal effects*. United States of America, Cambridge.

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

Fichman, M., & Cummings, J. N. (2003). Multiple imputation for missing data: Making the most of what you know. *Organizational research methods, 6(3)*, 282-308

Haberman, S. J. (2008). When can subscores have value? *Journal of educational and behavioral statistics, 33(2),* 204-229.

Haberman, S. J., & Sinharay, S. (2008). *Subscores based multidimensional item response theory*. Princeton, NJ: Educational Testing Services.

Han, K. J., & Hambleton, R. (2007) *WINGEN2: Windows software that generates IRT model parameters and item responses*[Computer Software]. Amherst, MA.

Holland, P. W., Sinharay, S., von Davier. A. A., & Han, N. (2007). An Approach to Evaluating the Missing Data Assumptions of the Chain and Post-stratification Equating Methods for the NEAT Design. *Journal of educational measurement, 45(1),* 17-43.

Holland, P. W., & Rubin, D. B. (1982). *Test equating.* New York: Academic Press

Holland, P. W., & Sinharay, S. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of educational measurement, 44(3),* 249-275.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of educational and behavioral statistics, 25(2),* 133-183.

Horton, N. J., Lipsitz, S. R., & Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American statisticians, 57(4),* 229-232.

Kolen, M. J. (1990). Does matching in equating works? A discussion. *Applied measurement in education, 3(1),* 97-104.

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

Luellen, J. K. (2007). *A comparison of propensity score estimation and adjustment methods on simulated data.* Unpublished doctoral dissertation, The University of Memphis, Memphis, Tennessee.

Livingston, S. A., Dorans, N.J., & Wright, N. K. (1992). What combination of sampling and equating methods works best? *Applied measurement in education, 3,* 73-95.

Little, R., & Rubin, D. B. & (2002). Statistical analysis with missing data. 2nd. New Jersey: John Wiley & Sons.

Lord, F. M. (1950). *Notes on comparable scales for test scores* (RB-50-48). Princeton, NJ: Educational Testing Service.

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

Mislevy, R. J., Kathleen, M., & Sheehan, K. M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika, 54,* 661-679.

Mostteller, F., Youtz, C. (1961). Tables of the Freeman-Tukey transformation for the binomial and Poison distributions. *Biometrika, 48,* 433-440.

Muthen, B., & Muthen, L. (2008). *Mplus 5.1* [Computer Software]. Los Angeles, CA: Muthen & Muthen

Paek, I., Liu, J., & Oh, H. (2008). *An investigation of propensity score matching on linear/nonlinear observed score equating method in a large scale assessment.* Paper presented at the Annual Meeting of National Council on Measurement in Education. 25- 27 March, New York.

Petersen, N. S. (2008). A Discussion of population invariance of equating. *Applied psychological of equating, 32(1),* 98-101.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New Jersey: Educational Testing Service.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of propensity score in observational studies for causal effects. *Biometrika, 70(1),* 41-55.

Roussos, L. A., Templin, J. L., & Henson, R. A. (2007). Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement, 44(4),* 293-312.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

SAS Institute. (2003). *SAS Stat.* Cary, NC: SAS Institute, Inc.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data,* New York: Chapman and Hall.

Schafer, J. L. & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological methods, 7(2),* 147-177.

Schmidt. W. H., et al (2007). *The preparation gaps: teacher education for middle school mathematics in six countries. (MT21 Report).* East Lansing, MI: Michigan State University.

Sinharay, S. & Haberman, S. & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational measurement: issues and practice, 26(4),* 21-28.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). *WinBUGS1.4* [Computer software].

Tate, R. L. (2004). Implications of multidimensionality for total score and subscore performance. *Applied measurement in education, 17(2),* 89-112.

Wang, T., Lee, W., Brennan, R. L., & Kolen, M. J. (2008). A comparison of the frequency estimation and chained equipercentile methods under the common-item non-equivalent groups design. *Applied Psychological Measurement,* 32(8), 632-651.

Wang, T., & Brennan, R. L. (2009). A modified frequency estimation equating method for the common-item non-equivalent groups. *Applied measurement in education, 33(2)*, 118-132.

Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (2002). *BILOG-MG3* [Computer Software]. Lincolnwood, IL: Scientific Software International, Inc.