Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 43

Equating with Bivariate Log-linear Presmoothing under the Common-item Nonequivalent Groups Design: Structural Zeros and Their Implications

Hyung Jin Kim[†]
Robert L. Brennan
Won-Chan Lee

June 2015

[†]Hyung Jin Kim is Associate Research Scientist, Center for Advanced Studies in Measurement and Assessment (CASMA), College of Education, University of Iowa (email: hyungjin-kim@uiowa.edu). Robert L. Brennan is E. F. Lindquist Chair in Measurement and Testing and Co-Director, Center for Advanced Studies in Measurement and Assessment (CASMA), College of Education, University of Iowa (email: robert-brennan@uiowa.edu). Won-Chan Lee is Co-Director, Center for Advanced Studies in Measurement and Assessment (CASMA), College of Education, University of Iowa (email: won-chan-lee@uiowa.edu).

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: www.education.uiowa.edu/casma

Contents

List of Tables

List of Figures

Abstract

In equating, when common items are internal and scoring is in terms of the number of correct items, some pairs of total scores (X) and common-item scores (V) can never be observed in a bivariate distribution of X and V; these pairs are called *structural zeros* (Bishop, Fienberg, & Holland, 2007; Holland & Wang, 1987). This study examines how different approaches to handling structural zeros give different equating results. It considers four different approaches: (1) no smoothing approach, (2) internal approach, (3) external approach, and (4) adjusted external approach. Operational and simulated test analyses led to four main findings: (1) the external approach generally gave the worst results; (2) for relatively small effect sizes, the internal approach generally gave the smallest overall error; (3) for relatively large effect sizes, the adjusted external approach generally had the smallest overall error; and (4) if sole interest focuses on reducing bias only, the adjusted external approach was generally preferable. These results suggest that, when a set of common items is an internal anchor and bivariate smoothing is performed, the internal approach which maintains structural zeros is generally preferable.

1 Introduction

Equating is a statistical process that adjusts for differences in difficulty among different forms of the same test so that scores on the forms can be used interchangeably (Kolen & Brennan, 2014). In order to improve the performance of equating, Rosenbaum and Thayer (1987) have suggested a process called presmoothing that performs log-linear smoothing on discrete score distributions before applying an equating procedure. When log-linear presmoothing is performed on score distributions, it is expected that fitted smoothed distributions maintain three properties. First, log-linear presmoothing should ensure that the first few moments of smoothed distributions are the same as those of observed distributions. Second, presmoothing should result in smoothed distributions. Third, presmoothing should assign an "appropriate" probability to each cell in the score distributions. Smoothing itself induces bias to some degree in equating results. Inability to maintain the first and the third properties also becomes an additional source of bias in equating results.

This study focuses on the common-item nonequivalent groups (CINEG) design, pays specific attention to bivariate smoothing with a set of common items as an internal anchor, and examines equating results using different approaches to smoothing score distributions. Principal focus is on structural zeros (discussed next), and their impact on equating error, particularly bias.

1.1 Definition of Structural Zeros

When a bivariate distribution is considered with rows and columns corresponding to two variables, it is sometimes impossible to observe pairs of variables, called *structural zeros* (Bishop et al., 2007; Holland & Wang, 1987). In other words, structural zeros refer to cells with zero probabilities of observing pairs of variables in the bivariate distribution. Form X, hereafter, refers to the new form administered later and Form Y refers to the old form administered earlier. Total scores on Form X and Form Y are represented by X and Y, respectively; and, scores on common items are represented by V. In order to understand the concept of structural zeros, consider Form X with internal common items. When the set of common items is an internal anchor, the total number of operational items ($K_X = K_V + K_U$) is the sum of the number of common items (K_V) and the number of non-common items (K_U). When scoring is conducted in terms of the number of correct items, the total score (X) for each examinee also becomes the sum of the score on common items (V) and the score on non-common items (U). For this internal case, the bivariate distribution of total score (X) and score on the common items (V) possesses structural zeros because some pairs of X and V can never be observed. For example, suppose that there are 8 items for a test form with 3 common items and 5 non-common items. Then, for an examinee with a perfect total score of 8, the score on the common items must be 3. Thus, the probability for an examinee obtaining a total score of 8 and a common-item score less than 3 is zero. For the given example, Table 1.1 presents all possible structural zeros in the bivariate $X \times V$ matrix. Note that, as a total score X increases to the highest score of 8 (or decreases to the lowest score of 0), it is associated with a greater number of structural zeros.

1.2 Issues and Potential Problems

As noted by Brennan, Wang, Kim, and Seol (2009),when common items are internal, bivariate smoothing on the $X \times V$ matrix can assign positive probabilities to cells with structural zeros in this matrix; i.e., structural zeros are ignored when bivariate smoothing is performed on the $X \times V$ matrix. In short, bivariate smoothing on the $X \times V$ matrix will not be able to maintain the third property when common items are internal. In order to solve this problem, when common items are an internal anchor, total scores (X) can be separated into scores on the common items

Table 1.1: Example of Structural Zeros

		Total Score (X)								
		0	1	2	3	4	5	6	7	8
Common	0	+	+	+	+	+	+	0	0	0
Item	1	0	+	+	+	+	+	+	0	0
Score	2	0	0	+	+	+	+	+	+	0
(V)	3	0	0	0	+	+	+	+	+	+

Note: 0 represents a combination where a structural zero occurs. + represents a combination where a positive probability occurs.

Table 1.2: Example of $U \times V$ Matrix

		Non-Common Item Score (U)					
		0	1	2	3	4	5
Common	0	+	+	+	+	+	+
Item	1	+	+	+	+	+	+
Score	2	+	+	+	+	+	+
(V)	3	+	+	+	+	+	+

Note: + represents a combination where a positive probability occurs.

(V) and scores on the non-common items (U) such that $X = V + U$ (Brennan et al., 2009). Using the example above, Table 1.2 shows that the $U \times V$ matrix does not have structural zeros. (i.e., a pair of variables can be observed in every cell). Smoothing can be also performed on the $U \times V$ matrix; and, these results can be used to build up the $X \times V$ matrix, such that structural zeros in the $X \times V$ matrix are maintained (i.e., probabilities for cells with structural zeros are all zero). Therefore, depending on how bivariate smoothing is performed, structural zeros are handled differently, either ignored or maintained.

According to the moment preservation property (Kolen & Brennan, 2014), log-linear presmoothing preserves moments of the smoothed marginal distributions that are the same as those of the observed marginal distributions. Therefore, when bivariate smoothing is performed on the $X \times V$ matrix, the moment preservation property is achieved for X and V. Similarly, when bivariate smoothing is performed on the $U \times V$ matrix, the moment preservation property is achieved for V and U, but not for X. In summary, this study focuses on two ways of performing bivariate log-linear smoothing; presmoothing the $X \times V$ matrix and presmoothing the $U \times V$ matrix. For presmoothing the $X \times V$ matrix, moments of X and V are preserved, but structural zeros are not maintained in the $X \times V$ matrix. For presmoothing the $U \times V$ matrix, although structural zeros are maintained in the built-up $X \times V$ matrix, the moment preservation property for X is lost, which might introduce bias in equating results. In short, when bivariate smoothing is performed on score distributions, the trade-off between maintaining structural zeros and ignoring structural zeros should be considered. For the purpose of convenience, "different approaches to handling structural zeros" and "different ways to performing bivariate smoothing" are used interchangeably here.

When the set of common items is an internal anchor, there are two important relationships between structural zeros and common items. First, the number of structural zeros equals $K_V(K_V + 1)$ where K_V represents the number of common items. Second, the proportion of structural zeros in the $X \times V$ matrix can be approximately estimated by the proportion of com-

mon items in the test. These two claims are proved in Appendices A and B. The second claim introduces problems when a larger proportion of common items is considered for equating.

According to Kolen and Brennan (2014), all other things being equal, a larger proportion of common items improves equating results through reducing equating error. However, an ambiguity (or contradiction) occurs when bivariate smoothing involves structural zeros. When smoothing is performed on the $X \times V$ matrix, assigning positive probabilities to structural zero cells becomes an additional source of bias. Thus, as the proportion of common items increases, the proportion of structural zeros also increases, which introduces more bias in equating results. Similarly, when smoothing is performed on the $U \times V$ matrix, the loss of the moment preservation property for X might increase as the proportion of structural zeros increases, which might also induce more bias in equating results. Therefore, in the context of equating with bivariate presmoothing, it cannot be stated unambiguously that equating is better with a larger proportion of common items.

This study investigates how different approaches to handling structural zeros (i.e., different ways to performing bivariate smoothing) give different equating results. Additionally, the study examines how the relationship between approaches to handling structural zeros and equating results changes as the proportion of common items changes.

1.3 Approaches to Handling Structural Zeros

Approaches to handling structural zeros depend on how smoothing is performed on bivariate distributions. This study considers five approaches to handling structural zeros: the no smoothing, internal, external, adjusted external, and univariate frequency estimation approaches. Each approach is expected to result in different degrees of bias introduced through the inability to both maintain structural zeros and preserve moments of the marginal (univariate) and bivariate distributions.

1.3.1 No Smoothing Approach

The no smoothing approach does not apply smoothing to either the $X \times V$ or the $U \times V$ matrix. Since the original bivariate distribution does not have structural zeros, the no smoothing approach maintains structural zeros. Moreover, since equating uses the original data, moments are perfectly preserved. However, the distributions for X and V are generally not smooth.

1.3.2 Internal Approach

For an internal anchor, total scores (X) can be separated into scores on the common items (V) and scores on the non-common items (U) (Brennan et al., 2009). The internal approach performs smoothing on the bivariate $U \times V$ distribution matrix, based on which the $X \times V$ matrix can be built. In doing so, moments for U and V are preserved; however, moments for X are not preserved. In other words, the internal approach maintains structural zeros, but loses the moment preservation property for X. Since smoothing is done on the $U \times V$ matrix, smoothness is achieved for the marginal and bivariate distributions of U and V. However, smoothness might not be achieved for the marginal distributions of X and the bivariate distribution of X and V.

1.3.3 External Approach

The external approach performs smoothing on the bivariate $X \times V$ distribution matrix (Brennan et al., 2009). As a result, the external approach maintains the moment preservation property of X and V. However, positive probabilities are assigned to cell with structural zeros; i.e.,

the external approach introduces bias in equating results from ignoring structural zeros. Since smoothing is performed on the $X \times V$ matrix, smoothness is achieved for the marginal and bivariate distributions of X and V.

1.3.4 Adjusted External Approach

The adjusted external approach also performs smoothing on the bivariate $X \times V$ matrix. However, unlike the external approach which ignores structural zeros, the adjusted external approach replaces those positive probabilities with zeros to maintain structural zeros. The remaining probabilities for non-structural zeros are, then, standardized through dividing each number by their sum; in doing so, the sum of probabilities equals one. Since adjustments are made for the probabilities, it is unknown how well the adjusted external approach preserves moments for X and V. Note that standardization is performed by dividing the smoothed probabilities by a constant. Thus, for the adjusted external approach, it is certain that smoothness is achieved for non-structural zeros; however, it is not certain whether or not smoothness is achieved for the marginal and the bivariate distributions of X and V.

1.3.5 Univariate Frequency Estimation Approach

The univariate frequency estimation (UFE) approach performs smoothing on the univariate (i.e., marginal) score distributions for the synthetic populations. First, based on the basic assumption of the unsmoothed frequency estimation method, the marginal distributions for total scores, X and Y, are obtained for the synthetic populations. Then, the UFE approach performs univariate log-linear smoothing on the marginal distributions of X and Y, followed by conducting equipercentile equating using the two smoothed marginal distributions.

The UFE approach is different from the other approaches with respect to when and what type of presmoothing is performed. While, for the internal, external, and adjusted external approaches, presmoothing is performed before synthetic populations are formed, the UFE approach performs presmoothing after the formation of the synthetic populations. Moreover, the UFE approach performs univariate log-linear presmoothing, whereas the internal, external, and adjusted external approaches perform bivariate log-linear presmoothing. Due to the differences in process, results for the UFE approach might not be directly comparable to results for the other approaches. However, since the marginal distributions do not have structural zeros, the UFE approach can avoid the issue regarding structural zeros.

2 Methodology

This section consists of two sections. The first section discusses how the operational test analyses were performed for five tests. The second section discusses the methodology implemented for the simulated test analyses.

2.1 Operational Test Analyses

Here, the first section discusses how operational test data sets were selected and prepared for analysis. The second section describes the study factors that were investigated. The last section explains how the adequacy of equating results was evaluated.

Table 2.1: Characteristics of Selected Tests

Test	Test Length	Proportion of Common Items	Sample Size (New, Old)	Effect Size
L1	85	23.53	(1170, 2271)	−0.3219
L2	85	24.71	(1819, 2478)	0.0919
M	50	32.00	(10708, 16417)	0.3207
S1	60	35.00	(7233, 11821)	0.1731
S2	75	29.33	(3557, 4782)	−0.0576

2.1.1 Data Preparation

Operational test data sets were collected under the CINEG design and items were dichotomously scored. Before tests were selected, scoring was conducted in terms of the number-correct items, and examinees who completed less than 80% of the items were removed. Five tests were selected so that they had differences with respect to content area, the proportion of common items, sample sizes, and common-item effect size.

The common-item effect size is the standardized mean difference in common-item scores between the new and old test forms, representing the ability difference between the new and old groups. The common-item effect size is calculated as follows:

$$\text{Effect Size} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(N_1-1)*s_1^2+(N_2-1)*s_2^2}{N_1+N_2}}} \tag{1}$$

where
$\bar{x}_1$ = Mean of common-item scores for the new form
$\bar{x}_2$ = Mean of common-item scores for the old form
N_1 = Number of examinees taking the new form
N_2 = Number of examinees taking the old form
s_1^2 = Variance of common-item scores based on common items for the new form
s_1^2 = Variance of common-item scores based on common items for the old form.

A positive value for the effect size indicates that, on average, examinees in the new group have a higher mean on the common items than examinees in the old group; a negative value represents the opposite case. Characteristics of tests are presented on Table 2.1.

2.1.2 Study Factors

For the operational test analyses, there were three study factors related to the choices of equating methods: approaches to handling structural zeros, equating methods, and the degrees of smoothing. For approaches to handling structural zeros, five different approaches were considered: the no smoothing, internal, external, adjusted external, and univariate frequency estimation (UFE) approaches. For equating methods, three different equipercentile equating methods were considered under the CINEG design: frequency estimation (FE), modified frequency estimation (MFE), and kernel equating methods. For kernel equating, the frequency estimation method was used for equating, which is, from now on, referred to as the kernel frequency estimation (KFE) method. These methods were selected because each method can involve presmoothing in its procedure. To simplify computations and interpretations, when a synthetic population was constructed, full

weight (one) was given to the new group taking Form X and zero weight to the old group taking Form Y.

Two different degrees of smoothing were considered in this study: 4 and 6. When bivariate smoothing was performed, the same degree of smoothing (4 or 6) was used for both X and Y (or U and V). A smoothing degree of 4 indicates that presmoothing preserves the first four moments of the marginal score distributions: mean, standard deviation, skewness, and kurtosis. Similarly, a smoothing degree of 6 indicates that the first six moments of the marginal score distributions are preserved. Additionally, for bivariate smoothing, the covariance between X and V (or U and V) was also preserved.

2.1.3 Evaluation

In order to evaluate equating results, the operational test analyses used unsmoothed frequency estimation with a one standard error band as the baseline for comparison. Since there are five operational test data sets, there are five different baseline equating relationships, one for each subject. Standard errors for equated raw scores were obtained using the nonparametric bootstrap method by taking multiple random samples with replacement; the number of replications was fixed at 500.

2.2 Simulated Test Analyses

The second study considers simulated test analyses. This section is divided into three subsections. The first section describes study factors investigated in the simulated test analyses. The second section discusses the procedure to simulate tests, followed by the last section explaining how equating results were evaluated.

2.2.1 Study Factors

In addition to the factors considered for the operational test analyses, the simulated test analyses considered four more study factors: test length, the proportion of common items, examinee ability effect size, and examinee sample size. Two different test lengths were considered: 30 items and 60 items. For the proportion of common items, three conditions were considered: 20%, 40%, and 60%. Three different sample sizes were considered for each group: 1,000, 3,000, and 6,000; thus, there were nine different pairs of sample sizes for the new and old groups of examinees.

Examinee ability effect size indicates the standardized mean difference in ability between the new and old groups. The examinee ability effect size is calculated as follows:

$$\text{Effect Size} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{(N_1-1)*\sigma_1^2+(N_2-1)*\sigma_2^2}{N_1+N_2}}} \tag{2}$$

where
μ_1 = Ability mean for the new group
μ_2 = Ability mean for the old group
N_1 = Number of examinees in the new group
N_2 = Number of examinees in the old group
σ_1^2 = Variance of abilities for examinees in the new group
σ_1^2 = Variance of abilities for examinees in the old group.

For the simulated test analyses, since population ability distributions for the new and old groups were known, true values for mean and standard deviation parameters could be used in

Table 2.2: True Ability Distributions for the New and Old Groups for Each Effect Size

Effect Size	New Group	Old Group
0.05	$N(0.05, 1)$	$N(0, 1)$
0.10	$N(0.10, 1)$	$N(0, 1)$
0.30	$N(0.30, 1)$	$N(0, 1)$
0.50	$N(0.50, 1)$	$N(0, 1)$

Equation 2; i.e., common-item scores were no longer needed to compute effect sizes. Four levels of the effect size were considered: 0.05, 0.10, 0.30, and 0.50. Under the assumption that ability follows the normal distribution, Table 2.2 provides population ability distributions for the new and old groups for each effect size. Population ability distributions were, then, used to sample examinees for the new and old groups.

For the factors of approaches to handling structural zeros, equating methods, and degree of smoothing, the same conditions as in the operational test analyses were considered for the simulated test analyses.

2.2.2 Simulation

In order to construct simulated tests, IRT item parameter estimates (a, b, and c) for the operational test S2 were considered. For the new and old forms of test S2, all item parameter estimates were obtained using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003). Item parameters for the simulated new form (Form X) were chosen from among item parameter estimates for the new form of test S2. Similarly, item parameters for the simulated old form (Form Y) were chosen from among item parameter estimates for the old form of test S2. Since Form X and Form Y should be parallel in statistical specifications, item parameters were carefully chosen so that the first two moments of item parameters were similar for both Form X and Form Y.

For each study condition considering the proportion of common items, test length, sample sizes, and effect sizes, item responses were generated using the 3PL IRT model for both the new group taking Form X and the old group taking Form Y. Scoring was conducted based on observed number-correct scores on total items and on common items. Thus, one set of simulated data consisted of number-correct scores on total items and on common items for both the new and old groups. Each simulated data set was used for performing all possible equating procedures which included all crossed factors of approaches to handling structural zeros, equating methods, and degrees of smoothing. For each study condition, one hundred data sets were simulated to compute random errors as well as systematic errors in equating results.

2.2.3 Evaluation

In order to evaluate equating results, the simulated test analyses used the IRT observed score equating method whose result was referred to as a population equating relationship. Since IRT was used to assemble simulated tests and generate data sets, it is reasonable to use an IRT equating method as a criterion. Moreover, when equating results are compared, the IRT observed score equating method does not give a particular equating method advantages over other methods. Since an IRT observed number-correct score distribution changes as the population ability distribution and test length change, there are 8 ($= 4 \times 2$) different population equating relationships based on study factors of test length and examinee ability effect size. When test length and examinee ability effect size were the same, the same population equating relationship

was used as the criterion regardless of sample sizes, approaches to handling structural zeros, equating methods, and the degrees of smoothing.

Equating results were evaluated in terms of both conditional statistics and overall statistics. For conditional statistics, bias (Bias), conditional standard error of equating (CSE), and root mean squared error (RMSE) can be calculated at each score point, x_i. Equations 3, 4, and 5 represent these statistics:

$$\text{Bias}_i = \frac{\sum_{j=1}^{J}(\hat{e}_Y(x_i) - e_Y(x_i))}{J} \tag{3}$$

$$\text{CSE}_i = \sqrt{\frac{\sum_{j=1}^{J}(\hat{e}_Y(x_i) - \bar{\hat{e}}_Y(x_i))^2}{J}} \tag{4}$$

$$\text{RMSE}_i = \sqrt{\text{Bias}_i^2 + \text{CSE}_i^2} \tag{5}$$

In Equations 3 to 5, i is a score point; x_i is a raw score at score point i; j is the j^{th} replication; J is the total number of replication; $e_Y(x_i)$ is Form Y equivalent of Form X score x_i for the criterion relationship; $\hat{e}_Y(x_i)$ is Form Y equivalent of Form X score x_i for an equating relationship under a study condition; and, $\bar{\hat{x}}_Y(x_i)$ is the average of Form Y equivalent of Form X score x_i for an equating relationship under a study condition over J replications. An index called "Difference That Matters" (DTM) was used to determine an acceptable level of bias at each score point for adequate equating. Since scoring was based on the number of correct items, the DTM value of 0.5 was used in this study.

In order to measure the overall amount of error over the entire score scale, both unweighted and weighted overall statistics were obtained. While unweighted statistics give equal weights to all score points, weighted statistics give weights equal to proportion of examinees scoring at each score point. Weighted average root mean squared bias (WRMSB), weighted average standard error of equating (WSE), and weighted average root mean squared error (WRMSE) can be calculated as follows:

$$\text{WRMSB} = \sqrt{\sum_i w_i \text{Bias}_i^2} \tag{6}$$

$$\text{WSE} = \sqrt{\sum_i w_i \text{CSE}_i^2} \tag{7}$$

$$\text{WRMSE} = \sqrt{\sum_i w_i \text{RMSE}_i^2} \tag{8}$$

For unweighted overall statistics, w_i in Equations 6 to 8 are replaced by $1/n$, where n is the number of raw score points. And, URMSB, USE, and URMSE are used to notate unweighted average root mean squared bias, unweighted average standard error of equating, and unweighted average root mean squared error, respectively.

3 Results

The first section provides results for the operational test analyses. The second section provides results for the simulated test analyses. For operational tests, when the degree of smoothing was 6, smoothed relative frequency distributions were closer to the observed relative frequency

distributions than when the degree of smoothing was 4. Similarly, equating relationships with 6 degrees of smoothing were closer to the baseline method (i.e., the unsmoothed frequency estimation equating method with the one standard error band) than those with 4 degrees of smoothing. Most importantly, for both operational test analyses and simulated test analyses, the statements and conclusions drawn are identical whether the degree of smoothing is 4 or 6. Therefore, results are provided only for 6 degrees of smoothing.

3.1 Operational Test Analyses

Results for the operational test analyses are presented with respect to three properties. The first and second sections compare smoothed relative frequency distributions and moments for different approaches to handling structural zeros. The third section presents equating relationships for different approaches to handling structural zeros.

3.1.1 Smoothed Relative Frequency Distributions

Figure 4.1 presents the observed relative frequency distributions for the five tests. Based on plot (a), the observed relative frequency distributions for test L1 were highly negatively skewed for both Form X_{L1} and Form Y_{L1}. Also, the observed relative frequency distributions for Form X_{L1} and Form Y_{L1} were quite different from each other. Based on plot (b) for test L2, the observed relative frequency distributions were also negatively skewed; but, the skewness was not as severe as the skewness observed for test L1. The observed relative frequency distributions were quite similar between the populations for Form X_{L2} and Form Y_{L2}. Figure 4.1 (c) shows that the observed relative frequency distributions for test M were quite smooth because of the large numbers of examinees for both Form X_M and Form Y_M. Based on plot (d) for test S1, the observed relative frequency distributions were negatively skewed for both Form X_{S1} and Form Y_{S1}. Additionally, they were quite smooth with large numbers of examinees, but not as smooth as those observed for test M. Based on plot (e), the observed relative frequency distributions for test S2 were not as smooth as those for test M and test S1. For Form X_{S2}, the observed relative frequency distribution was less negatively skewed compared to the other tests. Moreover, the observed relative frequency distributions for Form X_{S2} and Form Y_{S2} were quite different from each other.

Figure 4.2 contains two plots of the smoothed relative frequency distributions for test L1: plot (a) for Form X_{L1} and plot (b) for Form Y_{L1}. The vertical dashed line at a score of 65 represents the score above which structural zeros occurred in the bivariate score distributions. Based on plot (a), the largest discrepancies among different approaches occurred for scores with structural zeros, especially for very high scores and for scores with large relative frequencies. For high scores of 84 and 85, the relative frequencies for the external approach were somewhat higher than those for the internal and adjusted approaches. For scores of 76 to 83 with large relative frequencies, the relative frequencies for the internal approach were somewhat larger than those for the external and adjusted approaches. As scores became lower, smoothed relative frequencies became more similar across the internal, external, and adjusted approaches. Based on plot (b), similar results can be found for Form Y_{L1}; for Form Y_{L1}, discrepancies in relative frequencies were only larger than those for Form X_{L1}. Table 4.1 shows the average of absolute differences in relative frequencies for approaches relative to the observed relative frequencies, considering the whole score range as well as high scores where structural zeros occurred. Based on Table 4.1, for both Form X_{L1} and Form Y_{L1}, the external approach had smoothed relative frequencies closer, on average, to the observed relative frequencies than the internal and adjusted approaches.

For test L2, similar results were found for both Form X_{L2} and Form Y_{L2}. For high-end scores, the external approach gave somewhat higher relative frequencies than the internal and

adjusted approaches. For scores with high relative frequencies, relative frequencies for the internal approach were higher than those for the external and adjusted approaches. Overall, based on Table 4.1, the external approach gave smoothed relative frequencies closer to the observed relative frequencies than the other approaches. However, for test L2, the smoothed relative frequency distributions were much similar for different approaches than those for test L1. Note that the observed relative frequency distributions for test L2 were less negatively skewed than for test L1. Similar results were also found for test S1. For test S1, relative frequency distributions were even less negatively skewed than test L2; thus, smoothed relative frequency distributions were even similar for different approaches than those for test L2.

Based on Figure 4.4, for test M, the smoothed relative frequency distributions were all very similar for the different approaches to handling structural zeros. Note that, for test M, the observed relative frequency distributions were quite smooth since large numbers of examinees were involved. Therefore, smoothing the distribution did not add much more smoothness; and, the smoothed distributions were very similar among the internal, external, and adjusted approaches. In terms of discrepancies between observed and smoothed relative frequencies, Table 4.1 shows that the adjusted and external approaches gave smoothed relative frequencies closer to the observed values for Form X_M and Form Y_M, respectively.

Based on Figure 4.6 for test S2, the smoothed relative frequency distributions were all very similar for the different approaches, especially for Form X_{S2}. Note that, for Form X_{S2}, the observed relative frequency distribution was less negatively skewed. Consequently, the external approach assigned smaller positive probabilities to cells with structural zeros and the adjusted approach made less adjustment to the smoothed distribution. According to Table 4.1, the absolute differences were similar for all approaches.

The UFE approach gave smoothed relative frequencies that were substantially different from the other approaches. Note that the UFE approach performs smoothing on score distributions for synthetic populations. Since synthetic populations with $w_1 = 1$ were the same as the new population administered Form X, for Form X, the UFE approach gave a smoothed relative frequency distribution close to the observed distribution of raw scores X. However, for Form Y, the UFE approach gave a smoothed relative frequency distribution that was quite different from the observed Y distribution. For Form Y, discrepancies between observed and relative frequency distributions became smaller as the observed distributions for Form X and Form Y became more similar. Since the UFE approach gave smoothed relative frequency distributions for synthetic populations, those distributions should not be compared directly to the other smoothed distributions for the other approaches.

In summary, different approaches to handling structural zeros resulted in different smoothed relative frequency distributions, especially for scores where structural zeros occurred. For high-end scores where structural zeros occurred, the external approach gave somewhat larger smoothed relative frequencies than the internal and adjusted approaches. For scores where relative frequencies were high, the internal approach gave larger smoothed relative frequencies than the external and adjusted approaches. Differences became more evident when relative frequencies were higher for scores with structural zeros and when sample sizes were smaller. For example, differences were more evident for test L1 than for test S2. Observed score distributions for test S2 had less numbers of raw scores associated with structural zeros than those for test L1. With large sample sizes, different approaches gave smoothed relative frequency distributions that were similar to one another as well as the observed relative frequency distributions (i.e., test M). In terms of average of absolute differences in relative frequencies for approaches relative to those for the observed relative frequencies, Table 4.1 shows that, in general, the external approach gave smoothed relative frequency distributions that were the closest to the observed distributions.

3.1.2 Moment Preservation

Table 4.2 provides the differences between moments of the observed distribution and moments of the smoothed distribution using different approaches to handling structural zeros. Based on Table 4.2, the external approach preserved all moments. The internal approach preserved the first two moments, but it did not preserve the higher moments. Indeed, although the internal approach performs smoothing on the $U \times V$ matrix and preserves moments for U and V, it also preserves the first and second moments of X as long as the first cross-product moment, UV, as well as the first and second moments of V and U are preserved. The proof of this claim is provided in Appendix C.

The adjusted approach gave negative differences for the first and second moments; for the higher moments (i.e., the third to sixth moments), moments for the adjusted approach were closer to the observed moments than those for the internal approach. For the adjusted approach, the degree to which the first two moments were preserved depended on relative frequencies for scores with structural zeros, the numbers of examinees, and the proportions of common items. When relative frequencies were very high for scores with structural zeros (i.e., test L1), the adjusted approach made large adjustments to the relative frequencies of the smoothed distributions. Consequently, the adjusted approach gave large differences between the observed moments and the first two moments. By contrast, when relative frequencies were not high for scores with structural zeros (i.e., test S2), the adjusted approach tended to preserve the first two moments better. Moreover, when the observed relative frequency distributions had similar shapes, moments were better preserved with large number of examinees (i.e., test L2 vs. tests M and S1). When the shapes of relative frequency distributions were similar and sample sizes were large, the first two moments were better preserved with a smaller proportion of common items (i.e., test M vs. test S1). Note that a smaller proportion of common items is associated with a smaller proportion of structural zeros.

For the UFE approach, since $w_1 = 1$ for the synthetic populations, the synthetic population for Form X was the same as the new population administered Form X. Therefore, the UFE approach should give zero differences for all moments for Form X. However, since the synthetic population for Form Y was not the same as the old population administered Form Y, differences for all moments were substantial.

3.1.3 Equating Results

In general, compared to the FE and KFE methods, the MFE method gave the equating relationships that were the most different for different approaches to handling structural zeros. For high-end scores with structural zeros, the external approach with the MFE method resulted in equivalents that were noticeably lower than those using the other approaches; consequently, the MFE equivalents tended to be the furthest from the baseline method (i.e., the unsmoothed FE method). For the FE and KFE methods, the equating relationships were similar among the internal, external, and adjusted approaches relative to those for the MFE method. Figure 4.7 for test L1 presents these results the best.

Differences in equating results depended on the numbers of examinees and relative frequencies for scores with structural zeros. When the numbers of examinees were large, differences in equivalents were not large for different approaches to handling structural zeros. Note that, with large numbers of examinees, the smoothed relative frequency distributions were very similar among different approaches, which, in turn, led small differences in equivalents. For example, for tests M and S1, the differences in equivalents were smaller than a DTM of 0.5. Furthermore, based on Figures 4.7 and 4.8 for test L1 and test L2 respectively, when relative frequencies were high for scores with structural zeros, the differences in equivalents were larger across different

approaches with smaller numbers of examinees.

Although common items were an internal anchor for all tests, the internal approach did not always give an equating relationship that was closer to the unsmoothed FE method. In fact, if the 'unsmoothed' FE method is viewed as an ad hoc criterion, it is expected that the external approach will appear better than the internal approach because the external approach preserves the moments of total scores for both Form X and Form Y. Tables 4.3 through 4.7 present differences in unrounded equated scale scores for approaches relative to the internal approach for the five tests. For all tests, the largest differences in equivalents tended to occur at the low end of the score scale where there was very little data. In practice, however, equating is almost always conducted with a relatively small sample of the entire set of examinees who have taken the test. Consequently, it is reasonable that there will be substantial numbers of low scoring examinees in the population for whom equated scores must be repeated. Therefore, differences in equivalents at the low end of the scale have consequences.

3.2 Simulated Test Analyses

The simulated test analyses section is divided into eight subsections. The first section discusses conditional statistics for individual study conditions. The second section presents the main effect of approaches to handling structural zeros. The third to sixth sections present the interaction effect between approaches to handling structural zeros and the proportion of common items, test length, effect size, and sample size, respectively.

3.2.1 Conditional Statistics for Individual Study Conditions

Figure 4.12 contains three plots comparing bias among different approaches for the FE method: plot (a) for 20% CI, plot (b) for 40% CI, and plot (c) for 60% CI under the study conditions of a 60-item test with an effect size of 0.50 and sample sizes of 1000 for both the new and old groups. These study conditions were chosen because bias was larger and differences among different approaches were more evident relative to the other study conditions. Similarly, Figure 4.13 contains three plots comparing bias among different approaches for the MFE method. In each plot for Figures 4.12 and 4.13, a vertical dashed line represents the score above (or below) which structural zeros occur; for 60% CI, structural zeros can occur for all score points. For the FE method, when the proportion of common items was 20%, the internal approach gave smaller bias for high scores than the external approach. There were also scores for which the external approach gave smaller bias than the internal approach. As the proportion of common items increased, the score range for which the internal approach gave larger bias became wider. For 60% CI, bias for the internal approach seemed to increase even for low scores. For the MFE method, as the proportion of common items increased, the external approach introduced more bias for high scores where structural zeros occurred. However, for 60% CI, the internal approach resulted in larger bias for lower scores.

Figure 4.14 contains two plots comparing the CSEs among different approaches: plot (a) for the FE method and plot (b) for the MFE method under the study conditions of a 60-item test with 40% common items and an effect size of 0.50 with sample sizes of 1000 for both the new and old groups. Based on Figure 4.14, the internal approach gave somewhat smaller CSEs for most scores. Figure 4.15 contains two plots comparing the RMSEs among different approaches for the FE and MFE methods. Results for the RMSE were very similar to those found for bias discussed above.

3.2.2 Main Effect of Approaches to Handling Structural Zeros

Table 4.8 presents the main effect of approaches to handling structural zeros in terms of the unweighted and weighted overall statistics. For each equating method, overall statistics were compared among the internal, external, and adjusted approaches. Based on the unweighted overall statistics, for all equating methods, the adjusted approach gave the smallest URMSBs, whereas the internal approach gave the smallest USEs and URMSEs. For all equating methods, the largest URMSBs, USEs, and URMSEs resulted from using the external approach. Based on the weighted overall statistics, the adjusted approach gave the smallest WRMSBs and WRMSEs, whereas the internal approach gave the smallest WSEs. The external approach gave the largest WSEs and WRMSEs for all equating methods. However, with respect to the WRMSB, the largest values for the FE and KFE methods resulted from using the internal approach while the largest value for the MFE method resulted from using the external approach. The UFE approach tended to give overall statistics that were close to those for the external approach.

3.2.3 Interaction Effect between Proportion of Common Items and Approaches to Handling Structural Zeros

Tables 4.9 and 4.10 present the interaction between the proportion of common items and approaches to handling structural zeros in terms of the unweighted and weighted overall statistics, respectively. Within the same equating method and the same proportion of common items, each overall statistic was compared for the internal, external, and adjusted approaches. Based on Table 4.9, based on the USE and URMSE, results were similar to those found for the main effect of approaches to handling structural zeros; i.e., for all equating methods, the smallest and largest USEs (and URMSEs) resulted from using the internal and external approaches, respectively. However, results were quite different with respect to the URMSB. When the proportion of common items was 60%, the internal approach gave the largest URMSBs for the FE and KFE methods. For the MFE method, the internal approach gave the smallest URMSBs for 20% CI and 40% CI. With respect to the weighted overall statistics, Table 4.10 shows that results were similar to those found for the main effect of approaches to handling structural zeros.

Figure 4.16 presents nine plots showing the interaction between the proportion of common items and approaches to handling structural zeros in terms of URMSB, USE, and URMSE. Based on Figure 4.16, the unweighted overall statistics generally decreased as the proportion of common items increased. With respect to USE, the decreased amounts were similar for the same amount of increase in the proportion of common items; i.e., the decreases in the USE were similar between going from 20% to 40% and going from 40% to 60%. With respect to URMSB, results were quite different. As the proportion of common items increased from 40% to 60%, the decreases in URMSB were smaller than when the proportion of common items increased from 20% to 40%. Moreover, as the proportion of common items increased from 40% to 60%, the decreased amounts of bias were different for different approaches to handling structural zeros. The decreases for the internal approach seemed to be smaller than those for the other approaches. Results for the URMSE were similar to those for the URMSB. Figure 4.17 shows that results for the weighted overall statistics were similar to those for the unweighted overall statistics.

As noted previously, as the proportion of common items increases, the proportion of structural zeros also increases. Therefore, with a larger proportion of common items, the external approach is expected to induce more bias in equating results. For the internal approach, a different proportion of common items also introduces a different amount of bias associated with some degree of loss of the moment preservation property of X. However, unlike the external approach, for the internal approach, a larger proportion of common items does not always introduce more bias in equating results. Since total scores (X) consist of both common-item scores (V) and

scores on non-common items (U), the moment preservation property of X depends on which scores, V or U, comprise a larger proportion of X. Consider two proportions of common items, 10% and 20%; i.e., the proportions of non-common items are 90% and 80%, respectively. For these cases, since total scores (X) are determined mostly by scores on non-common items (U), the internal approach preserves moments of X better for 10% CI than for 20% CI. Note that the internal approach preserves moments of U and V. Consider another two proportions of common items, 80% and 90%; i.e., total scores (X) are determined mostly by scores on common items (V). Therefore, moments of X are preserved better for 90% CI than for 80% CI. As a result, it can be expected that moment preservation becomes worse as the proportion of common items increases up to a point; conversely, moment preservation becomes better as the proportion of common items increases beyond this point.

In order to confirm this assumption, one simulated dataset was selected for a 30-item test with an effect size of 0.50 and sample sizes of 6000 for both new and old groups. Nine different proportion of common items were considered for smoothing. Table 4.11 shows the differences between the actual moments and the moments for smoothed relative frequency distributions using the internal approach. Since the internal approach preserves the first two moments, differences were compared for the third to sixth moments. Based on Table 4.11, differences for the third moment increased until the proportion of common items reached 70% and started to decrease after then. Similar patterns were found for the other moments; i.e., differences increased and then decreased as the proportion of common items increased. Consequently, it can be expected that, for the internal approach, the smoothed relative frequency distribution will become less similar to the actual relative frequency distribution as the proportion of common items increases; i.e., the internal approach introduces more bias in equating results. However, above a certain proportion of common items, the smoothed relative frequency distribution becomes more similar to the actual relative frequency distribution; i.e., the internal approach preserves moments better and introduces less bias in equating results.

Based on the conditional statistics, as the proportion of common items increased, the bias was reduced for both the internal and external approaches. However, for 60% CI, the internal approach tended to give larger bias than the other approaches, even for low scores for which the internal approach gave smaller bias than the other approaches when the proportions of common items were 20% and 40%. This might explain why the decrease in bias for the internal approach was smaller than for the other approaches as the proportion of common items increased from 40% to 60%.

3.2.4 Interaction Effect between Test Length and Approaches to Handling Structural Zeros

Tables 4.12 and 4.13 present the interaction between test length and approaches to handling structural zeros using the unweighted and weighted overall statistics, respectively. Based on Table 4.12, results seemed to be consistent compared to those found for the main effect of approaches. Note that, for the MFE method, the smallest URMSBs for the 30-item test resulted from using the adjusted approach, whereas the smallest URMSBs for the 60-item test resulted from using the internal approach. Table 4.13 shows that results for the weighted overall statistics tended to be similar to those found for the main effect of approaches to handling structural zeros.

Figures 4.18 and 4.19 show that overall statistics were generally larger for the longer test. As the test length increased from 30 items to 60 items, the increase in overall statistics seemed to be very similar for different approaches. Recall that the IRT observed score equating method was used as the criterion. For the IRT observed score equating method, the population observed score distributions were smooth. With a fixed number of examinees, a relative frequency dis-

tribution is smoother for a shorter test than for a longer test. Therefore, a smoothed relative frequency distribution for the 30-item test is expected to be closer to the population observed score distribution than for the 60-item test. As a result, equating results for the 30-item test are expected to be more similar to the criterion relationship than those for the 60-item test, resulting in higher accuracy in equating results for the 30-item test.

3.2.5 Interaction Effect between Effect Size and Approaches to Handling Structural Zeros

Tables 4.14 and 4.15 present the interaction between effect size and approaches to handling structural zeros using the unweighted and weighted overall statistics, respectively. Based on Table 4.14, results for the unweighted overall statistics were similar to those found for the main effect of approaches regardless of effect sizes. Based on Table 4.15, results for the WRMSB and WSE were similar to those found for the main effect of approaches. However, results for the WRMSE were different between relatively small effect sizes and large effect sizes. For small effect sizes of 0.05 and 0.10, the internal approach gave the smallest WRMSEs; whereas, the adjusted approach gave the smallest WRMSEs for large effect sizes of 0.30 and 0.50. The largest WRMSEs resulted from using the external and internal approaches for small and large effect sizes, respectively. For the MFE method, the external approach gave the largest WRMSEs for all effect sizes. Overall, with respect to bias, the adjusted approach was preferable. If the focus was on reducing WRMSE, the adjusted approach was preferable for relatively large effect sizes (0.30 and 0.50) while the internal approach was preferable for relatively small effect sizes (0.05 and 0.10). Most importantly, for all effect sizes and equating methods, the external approach never gave the best results.

Figure 4.20 shows that the unweighted overall statistics increased as the effect size became larger. For the same amount of increase in the effect size, the increased amounts in the unweighted overall statistics were similar for different approaches to handling structural zeros; i.e., the interaction was not evident with respect the unweighted overall statistics. Similarly, based on Figure 4.21, the WSEs seemed to be very similar across different effect sizes. The WRMSBs and WRMSEs were larger for larger effect sizes. Moreover, the increases in the WRMSB and WRMSE seemed to be somewhat larger for the internal approach than for the other approaches.

3.2.6 Interaction Effect between Sample Size and Approaches to Handling Structural Zeros

Tables 4.16 and 4.17 present the interaction between sample size and approaches to handling structural zeros using the unweighted and weighted overall statistics, respectively. Tables 4.16 and 4.17 show that the results were similar to those found for the main effect of approaches to handling structural zeros, regardless of sample sizes. Based on Figure 4.22, the URMSBs were almost consistent across different sample sizes, whereas the USE decreased as sample sizes increased. Notice that, when the numbers of total examinees in both new and old groups were the same, the USEs were larger when the old group had a smaller sample size than the new group. For instance, when the total number of examinees was 4000, the USEs for (3000, 1000) were larger than those for (1000, 3000). Similar results were found for 7000 and 9000 for the total numbers of examinees (i.e., (6000, 1000) $\geq$ (1000, 6000) and (6000, 3000) $\geq$ (3000, 6000)). Results for the URMSE were similar to those for the USE. Figure 4.23 shows that the results for the weighted overall statistics were similar to those found for the unweighted overall statistics.

4 Discussion

This discussion session is divided into two subsections. The first subsection summarizes findings from the operational and simulated test analyses. The second subsection suggests future research based on limitations of this study.

4.1 Summary and Conclusions

Based on this study, although the UFE approach does not have to deal with structural zeros, it does not improve equating results; its performance was about as good as the external approach. For all approaches considered in this study, the performance of approaches to handling structural zeros depends on the proportion of common items. The increase in proportions of common items from 40% to 60% introduced more bias than the increase from 20% to 40%, especially for the internal approach. Moreover, the performance of approaches to handling structural zeros depends on the examinee ability effect size. In this study, for effect sizes of 0.05 and 0.10, the internal approach gave the smallest WRMSEs; however, for effect sizes of 0.30 and 0.50, the adjusted approach gave the smallest WRMSEs.

Overall, the results suggest that, if presmoothing is used with the CINEG design, the internal approach is almost always preferable to the external approach. Compared to the adjusted external approach, the internal approach is preferable when effect sizes are relatively small (0.05 and 0.10). For relatively large effect sizes (0.30 and 0.50), the adjusted external approach is preferable to the other approaches. If interest focuses on reducing bias only, the adjusted external approach is generally preferable. All things considered, these results suggest that bivariate smoothing should maintain structural zeros when the set of common items is an internal anchor.

It might be argued that differences among approaches were so small that any approach could be used to handle structural zeros. However, those small differences could have practical implications. As argued in the operational test analyses, small differences among approaches could lead to noticeable differences in rounded scale scores, which are the reported scores in most testing programs.[1] Those differences could become large as equating is to be performed over time. Therefore, even small differences among different approaches cannot be ignorable in a practical sense. On balance, the results suggest that, under the conditions studied here, the internal approach, which maintains structural zeros, is generally preferable when log-linear bivariate presmoothing is performed on bivariate distributions to conduct equating under the common-item nonequivalent groups design.

4.2 Limitations and Future Research

For the operational test analyses, without a criterion relationship, it is not permissible to conclude that any approach is better or worse; only similarities and differences among approaches have been assessed. Moreover, for many operational test forms, distributions of raw scores are negatively skewed. Therefore, future research is necessary to consider different observed score distributions.

For the simulated test analyses, the number of conditions was limited. For example, the kernel equating method considered only the use of the FE method, not the MFE method. Thus, for future research, the MFE method with kernel equating could also be considered. In addition, the continuized log-linear and chained equipercentile equating methods could be considered. Moreover, larger proportions of common items could also be considered in future research. As

[1] In fact, when replicating equating results for a testing program, reported scores were different depending on which approach, internal or external, was used; and, it caught the authors' attention.

the proportion of common items increases, the internal and external approaches introduce more bias. However, above a certain proportion of common items, bias introduced by the internal approach starts to decrease. Thus, it would be necessary to examine, in detail, patterns of bias introduced by the internal and external approaches as the proportion of common items increases.

For the simulated test analyses, the IRT observed score equating method was used as the criterion, expecting that it would not give a particular equating method an advantage over other methods. However, since this method employs smoothed distributions, it gave an advantage to shorter tests. That is, with other study conditions fixed, a smoothed relative frequency distribution was smoother for a shorter test length and closer to the population observed relative frequency distribution. Consequently, overall statistics were smaller for a shorter test than for a longer test. Therefore, future research could consider the use of a different criterion. One choice could be the chained equipercentile equating method with a large data set. Since marginal relative frequency distributions are usually smooth for a large data set, log-linear presmoothing may not be required for the chained equipercentile equating method.

Note also that pseudo-test forms (Hagge, 2010) could be used to examine how different results would be obtained for different approaches to handling structural zeros. With pseudo-test forms, the single group equipercentile method could be considered as a criterion. Then, future research could examine whether results are consistent regardless of how test forms are created.

Table 4.1: Average of Absolute Differences in Relative Frequencies for Approaches Compared to Those for the Observed Relative Frequencies

For High Scores with Structural Zeros

		Internal	External	Adjusted	UFE
L1	Form X	0.00440	0.00437	0.00456	0.00434
	Form Y	0.00471	0.00355	0.00467	0.01037
L2	Form X	0.00262	0.00250	0.00241	0.00250
	Form Y	0.00269	0.00247	0.00251	0.00265
M	Form X	0.00233	0.00194	0.00186	0.00190
	Form Y	0.00237	0.00191	0.00192	0.00694
S1	Form X	0.00237	0.00197	0.00200	0.00190
	Form Y	0.00199	0.00140	0.00153	0.00311
S2	Form X	0.00101	0.00102	0.00103	0.00101
	Form Y	0.00191	0.00186	0.00185	0.00353

For All Scores

		Internal	External	Adjusted	UFE
L1	Form X	0.00188	0.00184	0.00189	0.00183
	Form Y	0.00164	0.00133	0.00161	0.00341
L2	Form X	0.00148	0.00145	0.00143	0.00145
	Form Y	0.00132	0.00128	0.00130	0.00138
M	Form X	0.00126	0.00111	0.00108	0.00110
	Form Y	0.00114	0.00099	0.00100	0.00454
S1	Form X	0.00114	0.00101	0.00102	0.00098
	Form Y	0.00103	0.00082	0.00086	0.00211
S2	Form X	0.00099	0.00098	0.00098	0.00097
	Form Y	0.00109	0.00110	0.00109	0.00217

Table 4.2: Differences between Observed Moments and Moments Using Different Approaches to Handling Structural Zeros

		Form X				Form Y				
Test	Moment	I[a] - Obs*	E[b] - Obs	A[c] - Obs	U[d] - Obs	Moment	I - Obs	E - Obs	A - Obs	U - Obs
L1	1^{st}	(0)	(0)	-0.3789	(0)	1^{st}	(0)	(0)	-0.3303	-3.0845
	2^{nd}	(0)	(0)	-0.0498	(0)	2^{nd}	(0)	(0)	-0.0156	1.3639
	3^{rd}	-0.0409	(0)	-0.0027	(0)	3^{rd}	-0.0679	(0)	0.0070	0.1625
	4^{th}	0.0575	(0)	-0.0093	(0)	4^{th}	0.2217	(0)	-0.0213	-0.6080
	5^{th}	-0.2309	(0)	0.0486	(0)	5^{th}	-1.1018	(0)	0.1362	3.9289
	6^{th}	0.5394	(0)	-0.2918	(0)	6^{th}	5.1792	(0)	-0.7260	-21.7208
L2	1^{st}	(0)	(0)	-0.1068	(0)	1^{st}	(0)	(0)	-0.1089	1.0867
	2^{nd}	(0)	(0)	-0.0582	(0)	2^{nd}	(0)	(0)	-0.0633	-0.4268
	3^{rd}	-0.0495	(0)	-0.0100	(0)	3^{rd}	-0.0474	(0)	-0.0105	-0.0949
	4^{th}	0.0387	(0)	0.0051	(0)	4^{th}	0.0375	(0)	0.0036	0.2511
	5^{th}	-0.2869	(0)	-0.0712	(0)	5^{th}	-0.3233	(0)	-0.0720	-1.1809
	6^{th}	0.4297	(0)	0.1067	(0)	6^{th}	0.5502	(0)	0.0850	4.1608
M	1^{st}	(0)	(0)	-0.0409	(0)	1^{st}	(0)	(0)	-0.0611	2.5314
	2^{nd}	(0)	(0)	-0.0257	(0)	2^{nd}	(0)	(0)	-0.0316	-1.4182
	3^{rd}	-0.0329	(0)	-0.0148	(0)	3^{rd}	-0.0326	(0)	-0.0087	-0.1989
	4^{th}	0.0263	(0)	0.0191	(0)	4^{th}	0.0430	(0)	0.0039	0.8859
	5^{th}	-0.2795	(0)	-0.1582	(0)	5^{th}	-0.3128	(0)	-0.0560	-3.8357
	6^{th}	0.6656	(0)	0.4429	(0)	6^{th}	0.8172	(0)	0.0739	14.4918
S1	1^{st}	(0)	(0)	-0.0613	(0)	1^{st}	(0)	(0)	-0.1314	1.4923
	2^{nd}	(0)	(0)	-0.0329	(0)	2^{nd}	(0)	(0)	-0.0581	-1.1297
	3^{rd}	-0.0538	(0)	-0.0131	(0)	3^{rd}	-0.0422	(0)	-0.0141	-0.0394
	4^{th}	0.0628	(0)	0.0164	(0)	4^{th}	0.0446	(0)	0.0155	0.3471
	5^{th}	-0.4089	(0)	-0.1310	(0)	5^{th}	-0.3456	(0)	-0.1133	-1.5641
	6^{th}	1.0135	(0)	0.3533	(0)	6^{th}	0.9563	(0)	0.2658	7.0446
S2	1^{st}	(0)	(0)	-0.0256	(0)	1^{st}	(0)	(0)	-0.0571	-0.7648
	2^{nd}	(0)	(0)	-0.0255	(0)	2^{nd}	(0)	(0)	-0.0354	-1.0507
	3^{rd}	-0.0298	(0)	-0.0074	(0)	3^{rd}	-0.0369	(0)	-0.0070	0.0673
	4^{th}	-0.0184	(0)	-0.0050	(0)	4^{th}	0.0114	(0)	0.0025	0.1659
	5^{th}	-0.2036	(0)	-0.0608	(0)	5^{th}	-0.2285	(0)	-0.0497	0.0026
	6^{th}	-0.0534	(0)	-0.0360	(0)	6^{th}	0.3473	(0)	0.0557	1.8273

Note: (0) means 'must be zero'. Moments for the univariate frequency estimation approach are not directly comparable to moments for the other approaches.
$*$ Observed (actual) moments.
[a]Internal approach [b]External approach [c]Adjusted external approach dUnivariate frequency estimation approach

Table 4.3: Test L1: Differences in Unrounded Equated Scores for Approaches Relative to the Internal Approach

	FE				MFE			KFE	
Score	NS[e] - I[a]	U[d] - I	E[b] - I	A[c] - I	NS - I	E - I	A - I	E - I	A - I
0	**5.50**	**1.00**	0.50	0.50	**-2.50**	**1.00**	**1.00**	0.00	0.00
1	**5.50**	**1.00**	0.50	0.50	**-2.50**	**1.00**	**1.00**	0.00	0.00
2	**5.50**	**1.00**	0.50	0.50	**-2.50**	**1.00**	**1.00**	0.00	0.00
3	**5.50**	**1.00**	0.50	0.50	**-2.50**	**1.00**	**1.00**	0.00	0.00
4	**5.50**	**1.00**	0.50	0.50	**-2.50**	**1.00**	**1.00**	0.00	0.00
5	**5.50**	**1.00**	0.50	0.50	**-2.50**	**1.00**	**1.00**	0.00	0.00
6	**5.50**	**1.00**	0.50	0.50	**-2.50**	**1.00**	**1.00**	0.00	0.00
7	**5.50**	**1.00**	0.50	0.50	**-2.50**	**1.00**	**1.00**	0.00	0.00
8	**5.50**	**1.00**	0.50	0.50	**-2.50**	**1.00**	**1.00**	0.00	0.00
9	0.07	**1.87**	**1.14**	**1.14**	**-8.24**	**1.29**	**1.3**	0.00	0.00
10	**-0.96**	**1.80**	**1.10**	**1.10**	**-9.26**	**1.25**	**1.26**	0.00	0.00
11	**-2.00**	**1.73**	**1.06**	**1.06**	**-10.28**	**1.21**	**1.22**	0.00	0.00
12	**-3.06**	**1.64**	**1.03**	**1.03**	**-11.31**	**1.18**	**1.19**	**1.48**	**1.48**
13	**-4.13**	**1.54**	**0.98**	**0.98**	**-12.36**	**1.14**	**1.15**	**0.93**	**0.93**
14	**-5.23**	**1.42**	**0.93**	**0.93**	**-13.43**	**1.09**	**1.11**	**0.93**	**0.93**
15	**-6.35**	**1.29**	**0.88**	**0.88**	**-14.52**	**1.04**	**1.06**	**0.88**	**0.88**
16	**-7.50**	**1.13**	**0.82**	**0.82**	**-15.64**	**0.98**	**1.00**	**0.86**	**0.86**
17	**2.38**	**1.02**	**0.81**	**0.81**	**2.71**	**0.91**	**0.94**	**0.79**	**0.79**
18	**2.88**	**0.89**	**0.77**	**0.77**	**3.13**	**0.83**	**0.87**	**0.74**	**0.73**
19	**1.84**	**0.73**	**0.69**	**0.69**	**2.15**	**0.79**	**0.83**	**0.68**	**0.68**
20	**0.75**	**0.57**	**0.61**	**0.61**	**1.14**	**0.73**	**0.77**	**0.62**	**0.62**
21	-0.21	0.41	**0.53**	**0.53**	0.29	**0.64**	**0.68**	**0.55**	**0.55**
22	-0.12	0.28	0.49	0.49	0.16	**0.55**	**0.60**	0.48	0.48
23	**-1.24**	0.17	0.42	0.42	**-0.88**	0.48	**0.54**	0.42	0.42
24	**-2.40**	0.07	0.36	0.36	**-1.95**	0.42	0.48	0.36	0.37
25	**-1.08**	0.01	0.31	0.31	**-0.73**	0.35	0.41	0.31	0.31
26	**-2.02**	0.00	0.27	0.27	**-1.53**	0.30	0.37	0.27	0.28
27	-0.19	0.03	0.24	0.24	-0.01	0.26	0.33	0.24	0.24
28	**0.65**	0.10	0.22	0.22	**0.96**	0.22	0.30	0.22	0.22
29	-0.31	0.20	0.20	0.20	0.10	0.20	0.28	0.20	0.20
30	**0.67**	0.31	0.19	0.19	-0.07	0.18	0.26	0.19	0.19
31	**2.14**	0.41	0.18	0.18	**0.89**	0.16	0.25	0.18	0.18
32	**1.30**	0.49	0.17	0.17	**1.63**	0.15	0.24	0.17	0.17
33	**0.52**	**0.56**	0.16	0.16	**0.91**	0.14	0.23	0.16	0.16
34	0.13	**0.59**	0.15	0.15	0.47	0.12	0.20	0.14	0.14
35	**0.61**	**0.59**	0.13	0.13	**0.65**	0.10	0.19	0.13	0.13
36	0.41	**0.58**	0.12	0.12	**0.62**	0.09	0.17	0.11	0.11
37	0.25	**0.55**	0.10	0.10	0.11	0.08	0.16	0.10	0.10
38	-0.26	**0.52**	0.08	0.08	-0.13	0.07	0.14	0.08	0.08
39	**-0.80**	0.49	0.07	0.07	**-0.6**	0.06	0.13	0.07	0.07
40	**-0.53**	0.43	0.06	0.06	**-0.9**	0.05	0.13	0.06	0.05
41	0.06	0.38	0.04	0.04	**-0.7**	0.04	0.12	0.05	0.04
42	-0.29	0.32	0.04	0.04	-0.46	0.04	0.11	0.04	0.03
43	-0.19	0.28	0.03	0.03	-0.44	0.03	0.1	0.03	0.03
44	0.44	0.26	0.03	0.03	0.29	0.03	0.1	0.03	0.03
45	**0.88**	0.21	0.03	0.03	**0.78**	0.04	0.11	0.03	0.02
46	**1.10**	0.17	0.03	0.03	**1.09**	0.04	0.11	0.03	0.02
47	**1.15**	0.14	0.03	0.03	**1.20**	0.04	0.11	0.03	0.03

Table 4.3 Continued

	FE				MFE			KFE	
Score	NS[e] - I[a]	U[d] - I	E[b] - I	A[c] - I	NS - I	E - I	A - I	E - I	A - I
48	**0.85**	0.13	0.04	0.04	**0.95**	0.05	0.12	0.03	0.03
49	**0.64**	0.11	0.04	0.04	**0.68**	0.06	0.13	0.04	0.04
50	**0.84**	0.09	0.04	0.04	0.46	0.07	0.14	0.05	0.04
51	**0.51**	0.08	0.05	0.05	0.31	0.08	0.14	0.05	0.05
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
78	**-0.53**	-0.35	-0.05	-0.07	**-0.52**	-0.43	-0.04	-0.05	-0.07
79	**-0.74**	-0.37	-0.02	-0.06	**-0.80**	-0.45	-0.03	-0.02	-0.05
80	**-0.71**	-0.36	0.01	-0.04	**-0.88**	-0.47	-0.02	0.01	-0.04
81	-0.42	-0.33	0.04	-0.02	**-0.58**	-0.48	-0.01	0.04	-0.01
82	-0.13	-0.27	0.07	0.01	-0.32	-0.49	0.01	0.07	0.01
83	-0.05	-0.18	0.11	0.03	-0.19	-0.49	0.03	0.09	0.03
84	-0.01	-0.11	0.09	0.03	-0.09	-0.48	0.03	0.06	0.03
85	0.00	-0.08	-0.01	0.02	-0.02	-0.36	0.02	-0.01	0.03

Note: Boldface number refers to the difference greater than DTM of 0.5. Italic numbers refer to scores for which examinees were not observed.

[a]Internal approach [b]External approach [c]Adjusted external approach

[d]Univariate frequency estimation approach [e]No smoothing approach

Table 4.4: Test L2: Differences in Unrounded Equated Scores for Approaches Relative to the Internal Approach

	FE				MFE			KFE	
Score	NS[e] - I[a]	U[d] - I	E[b] - I	A[c] - I	NS - I	E - I	A - I	E - I	A - I
0	0.00	0.00	0.00	0.00	**0.50**	0.00	0.00	0.00	0.00
1	0.00	0.00	0.00	0.00	**0.50**	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	**0.50**	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	**0.50**	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	**0.50**	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	**0.50**	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	**0.50**	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	**0.50**	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00	**0.50**	0.00	0.00	0.00	0.00
9	**-1.08**	**-1.06**	**1.58**	**1.58**	**-0.30**	**0.76**	**0.76**	0.00	0.00
10	**-6.90**	**-6.80**	**3.15**	**3.15**	**-2.72**	**2.75**	**2.78**	0.00	0.00
11	**-11.22**	**-10.81**	**0.85**	**0.85**	**-9.86**	**1.21**	**1.22**	0.00	0.00
12	**-12.60**	**-11.24**	**0.71**	**0.71**	**-11.76**	**0.90**	**0.90**	**0.74**	**0.74**
13	**-13.89**	**-10.49**	**0.64**	**0.64**	**-13.04**	**0.84**	**0.84**	**0.56**	**0.56**
14	**-15.08**	**-8.65**	**0.68**	**0.68**	**-14.28**	**0.81**	**0.81**	**0.65**	**0.65**
15	**-16.24**	**-6.50**	**0.76**	**0.76**	**-15.51**	**0.79**	**0.79**	**0.70**	**0.7**
16	**-17.43**	**-4.62**	**0.73**	**0.73**	**-16.65**	**0.86**	**0.86**	**0.74**	**0.74**
17	**-18.65**	**-3.12**	**0.70**	**0.70**	**-17.83**	**0.83**	**0.83**	**0.70**	**0.70**
18	**0.75**	**-1.93**	**0.67**	**0.67**	**0.96**	**0.79**	**0.80**	**0.70**	**0.70**
19	**3.92**	**-1.01**	**0.67**	**0.67**	**3.62**	**0.75**	**0.75**	**0.68**	**0.68**
20	**3.03**	-0.28	**0.68**	**0.68**	**3.13**	**0.70**	**0.70**	**0.66**	**0.66**
21	**3.48**	0.24	**0.63**	**0.63**	**3.06**	**0.69**	**0.70**	**0.61**	**0.61**
22	**3.31**	**0.60**	**0.56**	**0.56**	**3.08**	**0.65**	**0.65**	**0.57**	**0.57**
23	**2.18**	**0.87**	0.50	0.50	**2.39**	**0.57**	**0.57**	**0.51**	**0.51**
24	**1.43**	**1.02**	0.45	0.45	**1.69**	0.48	0.49	0.45	0.45
25	**0.70**	**1.10**	0.39	0.39	**1.00**	0.41	0.41	0.38	0.38
26	-0.09	**1.12**	0.32	0.32	0.18	0.34	0.34	0.31	0.30
27	-0.19	**1.09**	0.23	0.23	-0.15	0.26	0.27	0.23	0.23
28	-0.26	**1.01**	0.15	0.15	-0.07	0.19	0.19	0.16	0.16
29	-0.32	**0.92**	0.08	0.08	-0.13	0.11	0.12	0.08	0.08
30	-0.11	**0.81**	0.02	0.02	-0.01	0.04	0.04	0.02	0.02
31	-0.41	**0.7**	-0.04	-0.04	-0.34	-0.02	-0.02	-0.04	-0.04
32	-0.25	**0.57**	-0.09	-0.09	-0.25	-0.07	-0.07	-0.09	-0.09
33	**0.51**	0.45	-0.13	-0.13	**0.53**	-0.11	-0.11	-0.13	-0.13
34	0.48	0.33	-0.16	-0.16	**0.54**	-0.15	-0.14	-0.16	-0.16
35	0.19	0.22	-0.18	-0.18	0.31	-0.17	-0.17	-0.18	-0.18
36	-0.06	0.12	-0.19	-0.19	-0.03	-0.19	-0.18	-0.19	-0.19
37	-0.13	0.04	-0.19	-0.19	-0.12	-0.20	-0.19	-0.20	-0.20
38	-0.23	-0.04	-0.19	-0.19	-0.21	-0.21	-0.20	-0.20	-0.20
39	-0.07	-0.10	-0.19	-0.19	-0.06	-0.20	-0.19	-0.19	-0.19
40	-0.03	-0.16	-0.19	-0.19	0.02	-0.19	-0.18	-0.18	-0.18
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
83	0.09	0.14	0.04	0.00	0.12	-0.43	-0.01	0.03	0.00
84	0.26	0.12	0.03	0.00	0.23	-0.5	-0.01	0.02	0.01
85	0.18	0.03	-0.01	0.01	0.16	-0.45	0.00	-0.01	0.01

Note: Boldface number refers to the difference greater than DTM of 0.5. Italic numbers refer to scores for which examinees were not observed.

[a]Internal approach [b]External approach [c]Adjusted external approach

[d]Univariate frequency estimation approach [e]No smoothing approach

Table 4.5: Test M: Differences in Unrounded Equated Scores for Approaches Relative to the Internal Approach

	FE				MFE			KFE	
Score	NS[e] - I[a]	U[d] - I	E[b] - I	A[c] - I	NS - I	E - I	A - I	E - I	A - I
0	**2.00**	0.50	**1.65**	0.50	**1.50**	**1.55**	0.50	**3.75**	**3.75**
1	0.45	**1.27**	**1.05**	**1.00**	**-0.14**	**0.89**	**0.68**	**3.75**	**3.75**
2	**-0.52**	**1.16**	**1.02**	**1.00**	**-1.12**	**0.73**	**0.61**	**0.94**	0.00
3	**-1.42**	**1.17**	**0.91**	**0.87**	**-2.10**	**0.55**	**0.51**	**0.94**	**0.94**
4	**1.34**	**1.17**	**0.80**	**0.79**	**0.74**	0.44	0.43	**0.82**	**0.82**
5	**0.83**	**0.95**	**0.73**	**0.73**	**0.54**	0.36	0.36	**0.73**	**0.73**
6	**0.83**	**0.82**	**0.67**	**0.67**	**0.64**	0.30	0.30	**0.65**	**0.65**
7	**0.62**	**0.72**	**0.61**	**0.61**	0.34	0.27	0.27	**0.55**	**0.55**
8	0.30	**0.62**	**0.51**	**0.50**	0.22	0.23	0.23	0.46	0.46
9	0.35	0.44	0.37	0.37	0.25	0.18	0.18	0.36	0.36
10	0.16	0.29	0.26	0.26	0.01	0.10	0.10	0.28	0.28
11	0.10	0.17	0.18	0.18	-0.10	0.03	0.03	0.19	0.19
12	0.17	0.06	0.10	0.10	-0.04	-0.03	-0.03	0.11	0.11
13	0.11	-0.05	0.04	0.04	-0.11	-0.07	-0.07	0.04	0.04
14	0.07	-0.15	-0.02	-0.02	-0.22	-0.10	-0.10	-0.01	-0.01
15	-0.22	-0.22	-0.07	-0.07	-0.38	-0.12	-0.12	-0.06	-0.06
16	-0.40	-0.27	-0.10	-0.10	-0.43	-0.12	-0.12	-0.09	-0.09
17	-0.49	-0.30	-0.12	-0.12	-0.49	-0.12	-0.12	-0.11	-0.11
18	-0.42	-0.30	-0.12	-0.12	-0.38	-0.10	-0.10	-0.12	-0.12
19	-0.24	-0.28	-0.12	-0.12	-0.25	-0.08	-0.08	-0.11	-0.11
20	-0.21	-0.24	-0.10	-0.10	-0.18	-0.06	-0.06	-0.10	-0.10
21	-0.08	-0.19	-0.08	-0.08	-0.06	-0.03	-0.03	-0.08	-0.08
22	0.01	-0.14	-0.05	-0.05	-0.04	-0.01	-0.01	-0.05	-0.05
23	-0.02	-0.09	-0.03	-0.03	0.04	0.00	0.00	-0.03	-0.03
24	0.05	-0.05	-0.01	-0.01	0.12	0.01	0.02	-0.01	-0.01
25	0.11	-0.02	0.01	0.01	0.11	0.02	0.02	0.01	0.01
26	0.10	0.00	0.02	0.02	0.09	0.02	0.02	0.02	0.02
27	0.07	0.01	0.03	0.03	0.09	0.02	0.02	0.02	0.02
28	0.00	0.00	0.03	0.02	0.01	0.01	0.01	0.02	0.02
29	-0.13	-0.01	0.02	0.02	-0.13	0.00	0.01	0.02	0.02
30	-0.25	-0.03	0.01	0.01	-0.25	-0.01	-0.01	0.01	0.01
31	-0.23	-0.04	0.00	0.00	-0.23	-0.02	-0.02	0.00	0.00
32	-0.14	-0.06	-0.01	-0.01	-0.14	-0.03	-0.03	-0.01	-0.01
33	-0.05	-0.07	-0.02	-0.02	-0.07	-0.04	-0.04	-0.02	-0.02
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
41	0.07	0.05	0.02	0.00	0.10	-0.02	0.00	0.02	0.00
42	0.12	0.08	0.04	0.01	0.18	-0.01	0.02	0.04	0.01
43	0.17	0.09	0.06	0.03	0.22	-0.01	0.03	0.06	0.03
44	0.15	0.10	0.08	0.04	0.17	-0.01	0.04	0.08	0.04
45	0.12	0.11	0.10	0.05	0.15	-0.03	0.04	0.09	0.04
46	0.13	0.12	0.12	0.06	0.15	-0.05	0.05	0.11	0.05
47	0.03	0.09	0.10	0.04	0.05	-0.07	0.03	0.11	0.04
48	0.05	0.08	0.10	0.02	0.06	-0.13	0.02	0.11	0.03
49	0.03	0.06	0.08	0.01	0.04	-0.23	0.00	0.09	0.01
50	-0.05	0.03	0.04	0.00	-0.04	-0.29	-0.01	0.06	0.00

Note: Boldface number refers to the difference greater than DTM of 0.5. Italic numbers refer to scores for which examinees were not observed.

[a]Internal approach [b]External approach [c]Adjusted external approach

[d]Univariate frequency estimation approach [e]No smoothing approach

Table 4.6: Test S1: Differences in Unrounded Equated Scores for Approaches Relative to the Internal Approach

	FE				MFE			KFE	
Score	NS[e] - I[a]	U[d] - I	E[b] - I	A[c] - I	NS - I	E - I	A - I	E - I	A - I
0	**2.50**	0.50	0.00	0.00	**2.00**	0.00	0.00	0.00	0.00
1	**2.50**	0.50	0.00	0.00	**2.00**	0.00	0.00	0.00	0.00
2	**2.50**	0.50	0.00	0.00	**2.00**	0.00	0.00	0.00	0.00
3	**2.50**	0.50	0.00	0.00	**2.00**	0.00	0.00	0.00	0.00
4	**2.50**	0.50	0.00	0.00	**2.00**	0.00	0.00	0.00	0.00
5	-0.10	**1.05**	-0.08	-0.06	**-0.66**	-0.14	-0.14	0.00	0.00
6	**-1.19**	**1.01**	-0.09	-0.08	**-1.78**	-0.18	-0.18	0.00	0.00
7	**-2.30**	**0.94**	-0.10	-0.09	**-2.93**	-0.20	-0.20	0.00	0.00
8	**1.71**	**0.85**	-0.09	-0.09	**1.44**	-0.20	-0.20	-0.07	-0.07
9	**1.49**	**0.75**	-0.07	-0.07	**1.27**	-0.19	-0.18	-0.03	-0.03
10	**0.67**	**0.63**	-0.05	-0.05	**0.50**	-0.15	-0.15	-0.03	-0.03
11	**1.09**	**0.51**	-0.03	-0.03	**0.90**	-0.08	-0.08	-0.03	-0.03
12	**0.53**	0.43	-0.01	-0.01	0.43	-0.07	-0.06	-0.01	-0.01
13	-0.16	0.36	0.00	0.00	-0.22	-0.06	-0.05	0.00	0.00
14	-0.14	0.27	0.00	0.00	-0.25	-0.04	-0.03	0.00	0.00
15	-0.18	0.19	0.00	0.00	-0.29	-0.03	-0.02	0.00	0.01
16	-0.06	0.10	0.00	0.00	0.01	-0.02	-0.01	0.00	0.00
17	0.01	0.03	0.00	0.00	-0.06	-0.02	-0.01	-0.01	0.00
18	0.02	-0.04	-0.02	-0.02	-0.06	-0.02	-0.01	-0.02	-0.01
19	0.01	-0.11	-0.03	-0.03	0.01	-0.03	-0.02	-0.03	-0.03
20	0.29	-0.16	-0.05	-0.05	0.28	-0.04	-0.03	-0.05	-0.04
21	0.25	-0.20	-0.07	-0.07	0.15	-0.05	-0.04	-0.06	-0.06
22	-0.19	-0.22	-0.08	-0.08	-0.25	-0.06	-0.05	-0.08	-0.08
23	-0.38	-0.24	-0.10	-0.10	-0.39	-0.08	-0.06	-0.10	-0.09
24	-0.34	-0.25	-0.11	-0.11	-0.40	-0.09	-0.07	-0.11	-0.10
25	**-0.51**	-0.24	-0.12	-0.12	**-0.53**	-0.09	-0.08	-0.11	-0.11
26	**-0.62**	-0.23	-0.12	-0.12	**-0.55**	-0.10	-0.08	-0.11	-0.11
27	-0.31	-0.21	-0.11	-0.11	-0.32	-0.10	-0.09	-0.11	-0.11
28	-0.12	-0.19	-0.10	-0.10	-0.11	-0.10	-0.09	-0.10	-0.10
29	0.04	-0.16	-0.09	-0.09	0.07	-0.09	-0.08	-0.09	-0.09
30	0.10	-0.14	-0.08	-0.08	0.09	-0.08	-0.06	-0.08	-0.07
31	-0.07	-0.11	-0.06	-0.06	-0.09	-0.06	-0.05	-0.06	-0.06
32	-0.12	-0.09	-0.05	-0.05	-0.12	-0.05	-0.04	-0.04	-0.04
33	-0.08	-0.07	-0.03	-0.03	-0.12	-0.03	-0.02	-0.03	-0.03
34	-0.11	-0.05	-0.01	-0.01	-0.12	-0.02	-0.01	-0.02	-0.01
35	-0.10	-0.04	0.00	0.00	-0.10	-0.01	0.00	0.00	0.00
36	-0.06	-0.03	0.01	0.01	**2.00**	0.00	0.01	0.01	0.01
37	0.04	-0.02	0.02	0.02	**2.00**	0.01	0.02	0.02	0.02
38	0.02	-0.02	0.02	0.02	-0.07	0.01	0.02	0.02	0.02
39	0.00	-0.02	0.02	0.02	0.04	0.01	0.02	0.02	0.02
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
57	0.11	0.08	0.13	-0.08	0.02	-0.28	-0.08	0.12	-0.08
58	0.08	0.11	0.16	-0.09	0.06	-0.32	-0.09	0.13	-0.08
59	-0.02	0.12	0.17	-0.04	0.11	-0.26	-0.04	0.08	-0.07
60	-0.02	-0.02	-0.01	-0.03	0.07	-0.24	-0.03	0.00	-0.05

Note: Boldface number refers to the difference greater than DTM of 0.5. Italic numbers refer to scores for which examinees were not observed.

[a]Internal approach [b]External approach [c]Adjusted external approach

[d]Univariate frequency estimation approach [e]No smoothing approach

Table 4.7: Test S2: Differences in Unrounded Equated Scores for Approaches Relative to the Internal Approach

	FE				MFE			KFE	
Score	NS[e] - I[a]	U[d] - I	E[b] - I	A[c] - I	NS - I	E - I	A - I	E - I	A - I
0	**2.50**	0.00	0.00	0.50	**2.50**	0.00	0.00	0.00	0.00
1	**2.50**	0.00	0.00	0.50	**2.50**	0.00	0.00	0.00	0.00
2	**2.50**	0.00	0.00	0.50	**2.50**	0.00	0.00	0.00	0.00
3	**2.50**	0.00	0.00	0.50	**2.50**	0.00	0.00	0.00	0.00
4	**2.50**	0.00	0.00	0.50	**2.50**	0.00	0.00	0.00	0.00
5	**-0.59**	**-3.09**	**-3.09**	**-2.59**	**-0.62**	**-3.12**	**-3.12**	0.00	0.00
6	**-1.57**	**-0.86**	-0.32	-0.32	**-1.59**	-0.40	-0.41	0.00	0.00
7	**-2.54**	**-0.81**	-0.34	-0.34	**-2.56**	-0.40	-0.41	**-0.66**	**-0.66**
8	**-3.50**	**-0.74**	-0.34	-0.34	**-3.53**	-0.39	-0.39	-0.33	-0.33
9	**-4.41**	**-0.62**	-0.27	-0.27	**-4.50**	-0.36	-0.36	-0.21	-0.21
10	**-5.35**	**-0.52**	-0.22	-0.22	**-5.44**	-0.30	-0.30	-0.19	-0.19
11	**-6.31**	-0.44	-0.17	-0.17	**-6.40**	-0.24	-0.24	-0.15	-0.15
12	**-0.79**	-0.37	-0.14	-0.14	**-0.79**	-0.20	-0.19	-0.11	-0.11
13	-0.40	-0.31	-0.11	-0.11	-0.45	-0.16	-0.16	-0.09	-0.09
14	**-0.98**	-0.26	-0.08	-0.08	**-1.06**	-0.13	-0.13	-0.08	-0.08
15	**-0.64**	-0.22	-0.07	-0.07	**-0.69**	-0.11	-0.10	-0.06	-0.06
16	-0.18	-0.18	-0.05	-0.05	-0.19	-0.07	-0.07	-0.05	-0.05
17	-0.24	-0.14	-0.04	-0.04	-0.22	-0.06	-0.06	-0.04	-0.04
18	-0.16	-0.11	-0.03	-0.03	-0.15	-0.06	-0.05	-0.04	-0.04
19	-0.15	-0.09	-0.03	-0.03	-0.19	-0.05	-0.05	-0.03	-0.03
20	-0.16	-0.07	-0.03	-0.03	-0.21	-0.05	-0.04	-0.03	-0.03
21	0.40	-0.05	-0.03	-0.03	**0.53**	-0.04	-0.03	-0.03	-0.03
22	**0.83**	-0.03	-0.02	-0.02	**0.82**	-0.02	-0.02	-0.02	-0.02
23	**0.93**	0.00	-0.01	-0.01	**0.91**	-0.01	0.00	-0.01	-0.01
24	**0.67**	0.03	0.01	0.01	**0.64**	0.01	0.02	0.01	0.01
25	0.33	0.05	0.02	0.02	0.29	0.03	0.04	0.02	0.02
26	-0.02	0.06	0.04	0.04	-0.05	0.05	0.06	0.04	0.04
27	-0.38	0.08	0.05	0.05	-0.39	0.06	0.07	0.05	0.05
28	**-0.56**	0.08	0.07	0.07	**-0.57**	0.07	0.08	0.07	0.06
29	**-0.71**	0.08	0.07	0.07	**-0.70**	0.08	0.09	0.07	0.07
30	-0.48	0.07	0.07	0.07	-0.47	0.07	0.08	0.07	0.07
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
49	-0.23	-0.07	-0.10	-0.11	-0.21	-0.11	-0.10	-0.10	-0.11
50	-0.31	-0.08	-0.11	-0.12	-0.30	-0.12	-0.11	-0.11	-0.12
51	-0.23	-0.10	-0.12	-0.13	-0.25	-0.13	-0.12	-0.12	-0.12
52	-0.20	-0.12	-0.13	-0.13	-0.20	-0.14	-0.13	-0.12	-0.13
53	-0.24	-0.14	-0.13	-0.14	-0.23	-0.15	-0.13	-0.13	-0.14
54	-0.32	-0.17	-0.13	-0.14	-0.30	-0.15	-0.14	-0.13	-0.14
55	-0.18	-0.19	-0.12	-0.14	-0.21	-0.15	-0.14	-0.12	-0.13
56	-0.09	-0.20	-0.11	-0.13	-0.13	-0.15	-0.13	-0.11	-0.13
57	-0.13	-0.21	-0.09	-0.11	-0.16	-0.14	-0.12	-0.10	-0.12
58	-0.18	-0.21	-0.07	-0.10	-0.19	-0.13	-0.10	-0.07	-0.10
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
74	-0.22	0.13	0.16	0.01	-0.25	-0.34	0.02	0.18	0.02
75	**-0.66**	0.06	0.07	0.00	**-0.65**	-0.46	0.01	0.11	0.01

Note: Boldface number refers to the difference greater than *DTM* of 0.5. Italic numbers refer to scores for which examinees were not observed.

[a]Internal approach [b]External approach [c]Adjusted external approach

[d]Univariate frequency estimation approach [e]No smoothing approach

Table 4.8: Main Effect of Approaches to Handling Structural Zeros

		Unweighted			Weighted		
		URMSB	USE	URMSE	WRMSB	WSE	WRMSE
FE	NS[e]	0.2898	0.2761	0.4310	0.3005	0.2334	0.4108
	UFE[d]	0.2878	0.2269	0.3930	0.2918	0.1844	0.3720
	I[a]	0.2840	**0.1973**	**0.3679**	0.3118	**0.1643**	0.3759
	E[b]	0.2896	0.2220	0.3900	0.3027	0.1810	0.3788
	A[c]	**0.2739**	0.2186	0.3763	**0.2810**	0.1771	**0.3589**
MFE	NS	0.1815	0.2759	0.3503	0.1857	0.2352	0.3195
	I	0.1811	**0.2016**	**0.2877**	0.2009	**0.1671**	0.2828
	E	0.2545	0.2222	0.3466	0.2605	0.1799	0.3303
	A	**0.1770**	0.2219	0.3026	**0.1798**	0.1798	**0.2764**
KFE	I	0.2844	**0.1954**	**0.3669**	0.3116	**0.1625**	0.3744
	E	0.2922	0.2193	0.3895	0.3029	0.1787	0.3772
	A	**0.2774**	0.2158	0.3760	**0.2820**	0.1748	**0.3576**

Note: Boldface number refers to the approach with the smallest value among the internal, external, and adjusted external approaches.

[a]Internal approach [b]External approach [c]Adjusted external approach

[d]Univariate frequency estimation approach [e]No smoothing approach

Table 4.9: Interaction between Proportion of Common Items and Approaches to Handling Structural Zeros for Equating Results (Unweighted Overall Statistics)

		URMSB			USE			URMSE		
		20% CI	40% CI	60% CI	20% CI	40% CI	60% CI	20% CI	40% CI	60% CI
FE	NS[e]	0.5127	0.2355	0.1211	0.3268	0.2749	0.2266	0.6427	0.3834	0.2670
	UFE[d]	0.5107	0.2335	0.1191	0.2851	0.2239	0.1718	0.6159	0.3441	0.2191
	I[a]	0.4951	0.2226	0.1342	**0.2687**	**0.1949**	**0.1283**	**0.5927**	**0.3144**	**0.1966**
	E[b]	0.5063	0.2375	0.1251	0.2838	0.2192	0.1630	0.6107	0.3427	0.2166
	A[c]	**0.4903**	**0.2223**	**0.1091**	0.2813	0.2160	0.1586	0.5959	0.3297	0.2032
MFE	NS	0.3201	0.1477	0.0766	0.3313	0.2735	0.2228	0.4881	0.3227	0.2402
	I	**0.2998**	**0.1380**	0.1056	**0.2787**	**0.1977**	**0.1283**	**0.4343**	**0.2535**	**0.1753**
	E	0.3675	0.2342	0.1619	0.2893	0.2182	0.1590	0.4854	0.3252	0.2292
	A	0.3142	0.1456	**0.0713**	0.2890	0.2181	0.1585	0.4529	0.2752	0.1797
KFE	I	0.4977	**0.2238**	0.1317	**0.2673**	**0.1924**	**0.1266**	**0.5935**	**0.3134**	**0.1938**
	E	0.5094	0.2396	0.1275	0.2820	0.2164	0.1594	0.6116	0.3417	0.2151
	A	**0.4935**	0.2254	**0.1132**	0.2793	0.2129	0.1551	0.5967	0.3290	0.2024

Table 4.10: Interaction between Proportion of Common Items and Approaches to Handling Structural Zeros for Equating Results (Weighted Overall Statistics)

		WRMSB			WSE			WRMSE		
		20% CI	40% CI	60% CI	20% CI	40% CI	60% CI	20% CI	40% CI	60% CI
FE	NS[e]	0.5130	0.2423	0.1462	0.2759	0.2292	0.1951	0.6172	0.3573	0.2578
	UFE[d]	0.5086	0.2337	0.1332	0.2356	0.1785	0.1390	0.5913	0.3169	0.2078
	I[a]	0.5227	0.2500	0.1626	**0.2237**	**0.1585**	**0.1106**	0.5982	0.3176	0.2118
	E[b]	0.5208	0.2463	0.1410	0.2342	0.1752	0.1335	0.6019	0.3248	0.2098
	A[c]	**0.4978**	**0.2246**	**0.1205**	0.2312	0.1711	0.1289	**0.5800**	**0.3051**	**0.1916**
MFE	NS	0.3050	0.1541	0.0979	0.2825	0.2304	0.1927	0.4438	0.2911	0.2236
	I	0.3196	0.1622	0.1209	**0.2300**	**0.1603**	**0.1110**	0.4233	0.2472	0.1778
	E	0.3751	0.2403	0.1660	0.2369	0.1729	0.1299	0.4666	0.3069	0.2174
	A	**0.3128**	**0.1476**	**0.0791**	0.2369	0.1729	0.1295	**0.4221**	**0.2452**	**0.1619**
KFE	I	0.5232	0.2498	0.1617	**0.2222**	**0.1565**	**0.1088**	0.5974	0.3159	0.2098
	E	0.5214	0.2462	0.1411	0.2325	0.1728	0.1309	0.6010	0.3228	0.2077
	A	**0.4985**	**0.2254**	**0.1222**	0.2293	0.1686	0.1264	**0.5791**	**0.3034**	**0.1903**

Note: Boldface number refers to the approach with the smallest value among the internal, external, and adjusted external approaches.
[a]Internal approach [b]External approach [c]Adjusted external approach [d]Univariate frequency estimation approach
[e]No smoothing approach

Table 4.11: An Example: Difference between Observed Moments and Moments for Smoothed Frequency Distributions with Different Proportion of Common Items Using the Internal Approach (30 Items, Effect Size 0.50, Sample Size (6000, 6000))

	Proportion of Common Items								
Moments	10%	20%	30%	40%	50%	60%	70%	80%	90%
3rd	-0.0439	-0.0665	-0.0702	-0.0751	-0.0849	-0.0895	-0.0903	-0.0459	-0.0108
4th	0.1115	0.1862	0.2107	0.1984	0.2115	0.1829	0.1754	0.1376	0.0394
5th	-0.5294	-0.8702	-1.0271	-0.9502	-1.0267	-0.8245	-0.7814	-0.5897	-0.1815
6th	1.8477	3.2315	4.0209	3.5025	3.7561	2.579	2.3176	2.1023	0.709

Table 4.12: Interaction between Test Length and Approaches to Handling Structural Zeros for Equating Results (Unweighted Overall Statistics)

		URMSB		USE		URMSE	
		30 Items	60 Items	30 Items	60 Items	30 Items	60 Items
FE	NS[e]	0.2586	0.3209	0.2208	0.3314	0.3639	0.4981
	UFE[d]	0.2571	0.3184	0.1884	0.2655	0.3399	0.4462
	I[a]	0.2514	0.3166	**0.1593**	**0.2353**	**0.3144**	**0.4215**
	E[b]	0.2558	0.3236	0.1859	0.2581	0.3364	0.4435
	A[c]	**0.2373**	**0.3104**	0.1836	0.2536	0.3216	0.4309
MFE	NS	0.1627	0.2002	0.2205	0.3313	0.2903	0.4104
	I	0.1666	**0.1957**	**0.1643**	**0.2388**	**0.2473**	**0.3281**
	E	0.2160	0.2931	0.1877	0.2566	0.2940	0.3992
	A	**0.1544**	0.1996	0.1874	0.2563	0.2585	0.3467
KFE	I	0.2523	0.3165	**0.1560**	**0.2349**	**0.3126**	**0.4212**
	E	0.2595	0.3248	0.1821	0.2564	0.3358	0.4431
	A	**0.2422**	**0.3124**	0.1792	0.2523	0.3210	0.4311

Table 4.13: Interaction between Test Length and Approaches to Handling Structural Zeros for Equating Results (Weighted Overall Statistics)

		WRMSB		WSE		WRMSE	
		30 Items	60 Items	30 Items	60 Items	30 Items	60 Items
FE	NS[e]	0.2723	0.3287	0.1669	0.2999	0.3422	0.4793
	UFE[d]	0.2698	0.3139	0.1413	0.2275	0.3246	0.4194
	I[a]	0.2827	0.3409	**0.1236**	**0.2049**	0.3250	0.4268
	E[b]	0.2751	0.3303	0.1395	0.2225	0.3281	0.4296
	A[c]	**0.2480**	**0.3139**	0.1364	0.2178	**0.3042**	**0.4136**
MFE	NS	0.1744	0.1970	0.1693	0.3011	0.2609	0.3781
	I	0.1935	0.2083	**0.1270**	**0.2072**	0.2479	0.3177
	E	0.2432	0.2778	0.1396	0.2202	0.2910	0.3695
	A	**0.1676**	**0.1920**	0.1396	0.2199	**0.2370**	**0.3157**
KFE	I	0.2828	0.3404	**0.1211**	**0.2039**	0.3231	0.4257
	E	0.2753	0.3304	0.1366	0.2208	0.3261	0.4283
	A	**0.2496**	**0.3144**	0.1332	0.2163	**0.3026**	**0.4126**

Note: Boldface number refers to the approach with the smallest value among the internal, external, and adjusted external approaches.

[a]Internal approach [b]External approach [c]Adjusted external approach

[d]Univariate frequency estimation approach [e]No smoothing approach

Table 4.14: Interaction between Effect Size and Approaches to Handling Structural Zeros for Equating Results (Unweighted Overall Statistics)

		URMSB				USE				URMSE			
		ES 0.05	ES 0.10	ES 0.30	ES 0.50	ES 0.05	ES 0.10	ES 0.30	ES 0.50	ES 0.05	ES 0.10	ES 0.30	ES 0.50
FE	NS[e]	0.0698	0.1229	0.3639	0.6025	0.2576	0.2632	0.2801	0.3035	0.2685	0.2955	0.4721	0.6880
	UFE[d]	0.0671	0.1208	0.3623	0.6008	0.2111	0.2160	0.2307	0.2500	0.2231	0.2519	0.4383	0.6588
	I[a]	0.0677	0.1215	0.3584	0.5883	**0.1850**	**0.1908**	**0.2005**	**0.2129**	**0.1984**	**0.2290**	**0.4151**	**0.6292**
	E[b]	0.0676	0.1234	0.3653	0.6023	0.2064	0.2123	0.2255	0.2439	0.2188	0.2493	0.4361	0.6558
	A[c]	**0.0592**	**0.1118**	**0.3470**	**0.5777**	0.2022	0.2083	0.2224	0.2415	0.2124	0.2405	0.4196	0.6327
MFE	NS	0.0498	0.0798	0.2255	0.3708	0.2589	0.2636	0.2793	0.3017	0.2643	0.2777	0.3685	0.4909
	I	0.0481	0.0787	0.2256	0.3722	**0.1898**	**0.1945**	**0.2048**	**0.2171**	**0.1967**	**0.2114**	**0.3083**	**0.4344**
	E	0.1353	0.1608	0.2916	0.4305	0.2068	0.2117	0.2256	0.2447	0.2491	0.2674	0.3712	0.4986
	A	**0.0431**	**0.0727**	**0.2213**	**0.3711**	0.2057	0.2110	0.2258	0.2450	0.2110	0.2254	0.3225	0.4516
KFE	I	0.0687	0.1221	0.3581	0.5886	**0.1828**	**0.1887**	**0.1985**	**0.2117**	**0.1969**	**0.2279**	**0.4138**	**0.6290**
	E	0.0704	0.1259	0.3672	0.6051	0.2029	0.2088	0.2231	0.2422	0.2166	0.2477	0.4362	0.6575
	A	**0.0636**	**0.1157**	**0.3495**	**0.5805**	0.1990	0.2049	0.2200	0.2391	0.2108	0.2393	0.4200	0.6340

Table 4.15: Interaction between Effect Size and Approaches to Handling Structural Zeros for Equating Results (Weighted Overall Statistics)

		WRMSB				WSE				WRMSE			
		ES 0.05	ES 0.10	ES 0.30	ES 0.50	ES 0.05	ES 0.10	ES 0.30	ES 0.50	ES 0.05	ES 0.10	ES 0.30	ES 0.50
FE	NS[e]	0.0822	0.1347	0.3749	0.6103	0.2348	0.2355	0.2323	0.2310	0.2503	0.2761	0.4531	0.6635
	UFE[d]	0.0673	0.1236	0.3700	0.6064	0.1856	0.1867	0.1833	0.1819	0.1995	0.2291	0.4204	0.6389
	I[a]	0.0700	0.1300	0.3928	0.6542	**0.1653**	**0.1674**	**0.1633**	**0.1609**	**0.1816**	**0.2158**	0.4295	0.6765
	E[b]	0.0689	0.1277	0.3819	0.6323	0.1822	0.1842	0.1799	0.1777	0.1968	0.2287	0.4284	0.6614
	A[c]	**0.0593**	**0.1140**	**0.3570**	**0.5937**	0.1783	0.1802	0.1760	0.1738	0.1899	0.2179	**0.4045**	**0.6234**
MFE	NS	0.0650	0.0911	0.2251	0.3615	0.2371	0.2368	0.2338	0.2332	0.2464	0.2554	0.3338	0.4424
	I	0.0470	0.0812	0.2490	0.4265	**0.1686**	**0.1699**	**0.1665**	**0.1634**	**0.1761**	**0.1906**	0.3039	0.4605
	E	0.1174	0.1493	0.3045	0.4708	0.1821	0.1830	0.1786	0.1759	0.2191	0.2390	0.3568	0.5063
	A	**0.0416**	**0.0714**	**0.2237**	**0.3825**	0.1812	0.1824	0.1789	0.1765	0.1868	0.1983	**0.2930**	**0.4275**
KFE	I	0.0715	0.1307	0.3914	0.6526	**0.1633**	**0.1652**	**0.1616**	**0.1599**	**0.1805**	**0.2147**	0.4275	0.6748
	E	0.0703	0.1283	0.3812	0.6318	0.1794	0.1812	0.1780	0.1764	0.1948	0.2268	0.4267	0.6605
	A	**0.0623**	**0.1159**	**0.3569**	**0.5931**	0.1757	0.1773	0.1739	0.1720	0.1885	0.2165	**0.4033**	**0.6222**

Note: Boldface number refers to the approach with the smallest value among the internal, external, and adjusted external approaches.
[a]Internal approach [b]External approach [c]Adjusted external approach [d]Univariate frequency estimation approach
[e]No smoothing approach

Table 4.16: Interaction between Sample Size and Approaches to Handling Structural Zeros for Equating Results (Unweighted Overall Statistics)

(a) URMSB

		Sample Size (New, Old)								
		(1k, 1k)	(1k, 3k)	(1k, 6k)	(3k, 1k)	(3k, 3k)	(3k, 6k)	(6k, 1k)	(6k, 3k)	(6k, 6k)
FE	NS[e]	0.2913	0.2949	0.2960	0.2903	0.2869	0.2871	0.2868	0.2873	0.2872
	UFE[d]	0.2769	0.2787	0.2795	0.2785	0.2729	0.2723	0.2747	0.2765	0.2747
	I[a]	0.2909	0.2906	0.2912	0.2873	0.2800	0.2792	0.2800	0.2780	0.2787
	E[b]	0.2929	0.2925	0.2923	0.2933	0.2873	0.2864	0.2863	0.2884	0.2875
	A[c]	**0.2755**	**0.2765**	**0.2762**	**0.2772**	**0.2718**	**0.2709**	**0.2706**	**0.2736**	**0.2728**
MFE	NS	0.1834	0.1830	0.1801	0.1864	0.1776	0.1771	0.1844	0.1817	0.1796
	I	0.1859	0.1840	0.1838	0.1854	0.1767	0.1771	0.1813	**0.1775**	0.1787
	E	0.2597	0.2577	0.2571	0.2572	0.2514	0.2517	0.2516	0.2523	0.2524
	A	**0.1772**	**0.1766**	**0.1751**	**0.1817**	**0.1741**	**0.1741**	**0.1766**	0.1795	**0.1785**
KFE	I	0.2911	0.2912	0.2918	0.2874	0.2808	0.2798	0.2799	**0.2786**	0.2791
	E	0.2937	0.2939	0.2938	0.2952	0.2903	0.2895	0.2891	0.2923	0.2915
	A	**0.2774**	**0.2788**	**0.2785**	**0.2800**	**0.2758**	**0.2750**	**0.2742**	0.2786	**0.2776**

(b) USE

		Sample Size (New, Old)								
		(1k, 1k)	(1k, 3k)	(1k, 6k)	(3k, 1k)	(3k, 3k)	(3k, 6k)	(6k, 1k)	(6k, 3k)	(6k, 6k)
FE	NS[e]	0.3536	0.2886	0.2751	0.3359	0.2432	0.2116	0.3521	0.2322	0.1928
	UFE[d]	0.2418	0.2034	0.1946	0.2326	0.1710	0.1509	0.2360	0.1622	0.1345
	I[a]	**0.2589**	**0.2147**	**0.2064**	**0.2373**	**0.1732**	**0.1525**	**0.2369**	**0.1618**	**0.1338**
	E[b]	0.2831	0.2363	0.2270	0.2648	0.1958	0.1733	0.2721	0.1879	0.1579
	A[c]	0.2779	0.2319	0.2232	0.2610	0.1931	0.1711	0.2679	0.1853	0.1559
MFE	NS	0.3544	0.2908	0.2792	0.3331	0.2433	0.2127	0.3471	0.2301	0.1923
	I	**0.2651**	**0.2198**	**0.2136**	**0.2414**	**0.1768**	**0.1555**	**0.2415**	**0.1641**	**0.1362**
	E	0.2836	0.2376	0.2309	0.2633	0.1959	0.1742	0.2696	0.1866	0.1580
	A	0.2829	0.2362	0.2293	0.2638	0.1957	0.1738	0.2703	0.1867	0.1579
KFE	I	**0.2570**	**0.2141**	**0.2057**	**0.2351**	**0.1710**	**0.1505**	**0.2347**	**0.1594**	**0.1315**
	E	0.2804	0.2341	0.2248	0.2622	0.1929	0.1707	0.2687	0.1847	0.1547
	A	0.2756	0.2299	0.2213	0.2581	0.1902	0.1683	0.2643	0.1818	0.1524

Note: Boldface number refers to the approach with the smallest value among the internal, external, and adjusted external approaches.
[a]Internal approach [b]External approach [c]Adjusted external approach [d]Univariate frequency estimation approach
[e]No smoothing approach

Table 4.16 Continued

(c) URMSE

		Sample Size (New, Old)								
		(1k, 1k)	(1k, 3k)	(1k, 6k)	(3k, 1k)	(3k, 3k)	(3k, 6k)	(6k, 1k)	(6k, 3k)	(6k, 6k)
FE	NS[e]	0.4912	0.4434	0.4325	0.4760	0.4044	0.3807	0.4847	0.3967	0.3696
	UFE[d]	0.3928	0.3681	0.3617	0.3881	0.3427	0.3285	0.3872	0.3400	0.3221
	I[a]	**0.4155**	**0.3843**	**0.3792**	**0.3963**	**0.3488**	**0.3341**	**0.3898**	**0.3398**	**0.3235**
	E[b]	0.4351	0.4019	0.3943	0.4219	0.3697	0.3533	0.4219	0.3658	0.3461
	A[c]	0.4197	0.3873	0.3798	0.4083	0.3565	0.3400	0.4084	0.3531	0.3332
MFE	NS	0.4175	0.3628	0.3503	0.4013	0.3211	0.2947	0.4113	0.3126	0.2816
	I	**0.3404**	**0.3032**	**0.2981**	**0.3212**	**0.2660**	**0.2492**	**0.3175**	**0.2564**	**0.2374**
	E	0.3926	0.3586	0.3527	0.3774	0.3256	0.3113	0.3772	0.3207	0.3031
	A	0.3511	0.3132	0.3057	0.3393	0.2799	0.2619	0.3413	0.2767	0.2544
KFE	I	**0.4141**	**0.3844**	**0.3791**	**0.3948**	**0.3476**	**0.3332**	**0.3883**	**0.3386**	**0.3222**
	E	0.4336	0.4011	0.3936	0.4211	0.3693	0.3533	0.4209	0.3659	0.3464
	A	0.4188	0.3869	0.3797	0.4076	0.3563	0.3403	0.4076	0.3533	0.3337

Note: Boldface number refers to the approach with the smallest value among the internal, external, and adjusted external approaches.
[a]Internal approach [b]External approach [c]Adjusted external approach [d]Univariate frequency estimation approach
[e]No smoothing approach

Table 4.17: Interaction between Sample Size and Approaches to Handling Structural Zeros for Equating Results (Weighted Overall Statistics)

(a) WRMSB

		Sample Size (New, Old)								
		(1k, 1k)	(1k, 3k)	(1k, 6k)	(3k, 1k)	(3k, 3k)	(3k, 6k)	(6k, 1k)	(6k, 3k)	(6k, 6k)
FE	NS[e]	0.3110	0.3099	0.3159	0.3011	0.2949	0.2935	0.2933	0.2919	0.2931
	UFE[d]	0.2924	0.2927	0.2963	0.2948	0.2902	0.2893	0.2900	0.2896	0.2912
	I[a]	0.3126	0.3111	0.3134	0.3168	0.3103	0.3093	0.3115	0.3095	0.3113
	E[b]	0.3047	0.3031	0.3058	0.3074	0.3009	0.2994	0.3020	0.3000	0.3011
	A[c]	**0.2828**	**0.2824**	**0.2851**	**0.2848**	**0.2789**	**0.2774**	**0.2796**	**0.2784**	**0.2794**
MFE	NS	0.2042	0.2009	0.2043	0.1868	0.1771	0.1750	0.1771	0.1729	0.1728
	I	0.2032	0.2004	0.2024	0.2056	0.1986	0.1985	0.2007	0.1984	0.2004
	E	0.2622	0.2593	0.2618	0.2643	0.2591	0.2587	0.2602	0.2590	0.2599
	A	**0.1834**	**0.1822**	**0.1843**	**0.1836**	**0.1765**	**0.1756**	**0.1782**	**0.1768**	**0.1777**
KFE	I	0.3122	0.3110	0.3134	0.3165	0.3102	0.3092	0.3112	0.3093	0.3111
	E	0.3048	0.3035	0.3061	0.3073	0.3011	0.2996	0.3019	0.3002	0.3014
	A	**0.2838**	**0.2835**	**0.2861**	**0.2855**	**0.2800**	**0.2786**	**0.2804**	**0.2796**	**0.2806**

(b) WSE

		Sample Size (New, Old)								
		(1k, 1k)	(1k, 3k)	(1k, 6k)	(3k, 1k)	(3k, 3k)	(3k, 6k)	(6k, 1k)	(6k, 3k)	(6k, 6k)
FE	NS[e]	0.3465	0.2851	0.2719	0.2735	0.1954	0.1700	0.2534	0.1670	0.1378
	UFE[d]	0.2727	0.2213	0.2097	0.2198	0.1548	0.1337	0.2046	0.1332	0.1097
	I[a]	**0.2413**	**0.1968**	**0.1875**	**0.1943**	**0.1380**	**0.1201**	**0.1825**	**0.1197**	**0.0980**
	E[b]	0.2661	0.2187	0.2079	0.2131	0.1525	0.1321	0.1991	0.1312	0.1083
	A[c]	0.2603	0.2140	0.2039	0.2085	0.1493	0.1294	0.1943	0.1281	0.1058
MFE	NS	0.3479	0.2891	0.2772	0.2736	0.1970	0.1722	0.2537	0.1673	0.1390
	I	**0.2447**	**0.1996**	**0.1919**	**0.1980**	**0.1403**	**0.1220**	**0.1865**	**0.1212**	**0.0997**
	E	0.2638	0.2178	0.2091	0.2115	0.1514	0.1315	0.1973	0.1292	0.1075
	A	0.2634	0.2167	0.2078	0.2121	0.1514	0.1312	0.1981	0.1295	0.1074
KFE	I	**0.2392**	**0.1956**	**0.1864**	**0.1921**	**0.1362**	**0.1186**	**0.1802**	**0.1178**	**0.0965**
	E	0.2636	0.2167	0.2057	0.2106	0.1502	0.1302	0.1962	0.1290	0.1064
	A	0.2577	0.2120	0.2017	0.2058	0.1470	0.1274	0.1912	0.1258	0.1038

Note: Boldface number refers to the approach with the smallest value among the internal, external, and adjusted external approaches.
[a]Internal approach [b]External approach [c]Adjusted external approach [d]Univariate frequency estimation approach
[e]No smoothing approach

Table 4.17 Continued

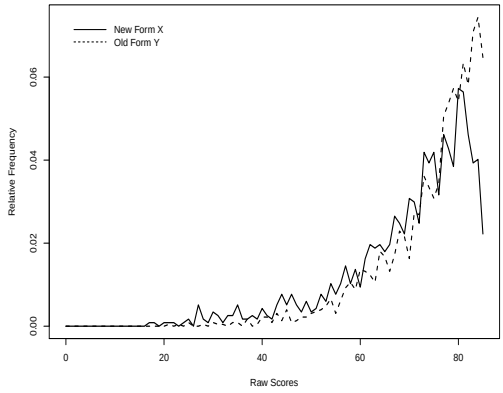
(c) WRMSE

		Sample Size (New, Old)								
		(1k, 1k)	(1k, 3k)	(1k, 6k)	(3k, 1k)	(3k, 3k)	(3k, 6k)	(6k, 1k)	(6k, 3k)	(6k, 6k)
FE	NS[e]	0.4971	0.4500	0.4434	0.4390	0.3808	0.3623	0.4189	0.3614	0.3439
	UFE[d]	0.4317	0.3952	0.3892	0.3967	0.3514	0.3369	0.3820	0.3384	0.3264
	I[a]	0.4248	0.3935	0.3896	0.3971	0.3586	0.3471	0.3849	0.3485	0.3386
	E[b]	0.4360	0.4016	0.3958	0.4022	0.3592	0.3450	0.3882	0.3468	0.3348
	A[c]	**0.4160**	**0.3824**	**0.3768**	**0.3820**	**0.3392**	**0.3250**	**0.3678**	**0.3267**	**0.3145**
MFE	NS	0.4194	0.3680	0.3594	0.3517	0.2843	0.2630	0.3299	0.2605	0.2392
	I	**0.3403**	0.3041	0.2998	0.3074	0.2621	0.2489	0.2950	0.2497	0.2377
	E	0.3867	0.3523	0.3477	0.3525	0.3103	0.2980	0.3391	0.2981	0.2877
	A	0.3422	**0.3041**	**0.2979**	**0.3032**	**0.2531**	**0.2371**	**0.2881**	**0.2382**	**0.2237**
KFE	I	0.4228	0.3924	0.3885	0.3951	0.3571	0.3459	0.3830	0.3471	0.3374
	E	0.4340	0.4001	0.3942	0.4002	0.3575	0.3437	0.3860	0.3453	0.3336
	A	**0.4142**	**0.3812**	**0.3755**	**0.3801**	**0.3380**	**0.3241**	**0.3658**	**0.3257**	**0.3137**

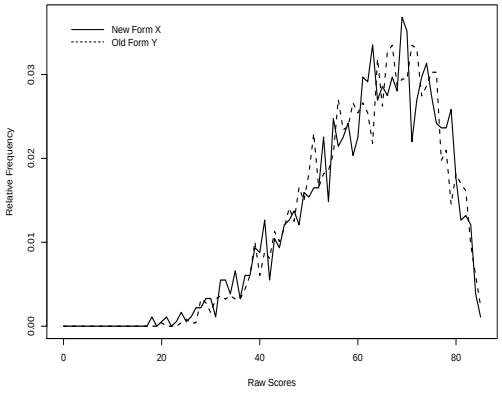
Note: Boldface number refers to the approach with the smallest value among the internal, external, and adjusted external approaches.
[a]Internal approach [b]External approach [c]Adjusted external approach [d]Univariate frequency estimation approach
[e]No smoothing approach

Figure 4.1: Observed Relative Frequency Distributions for Operational Tests

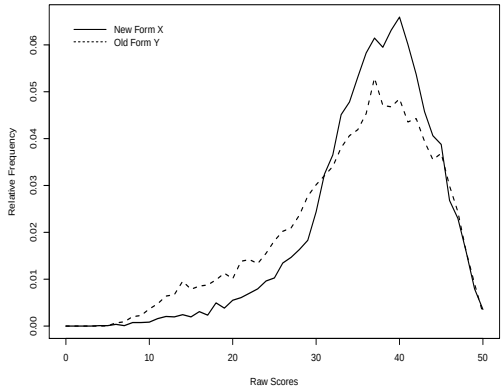
(a) Test L1



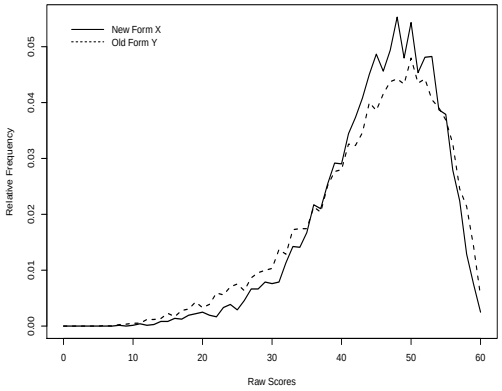
(b) Test L2



(c) Test M



(d) Test S1



(e) Test S2

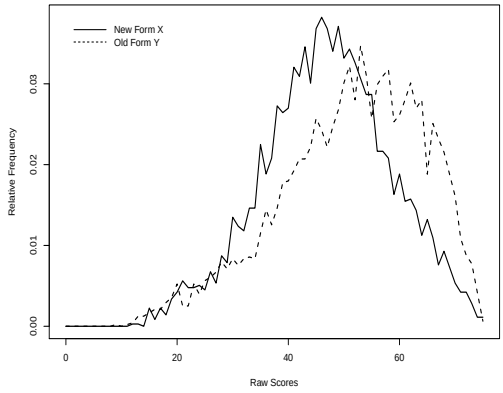
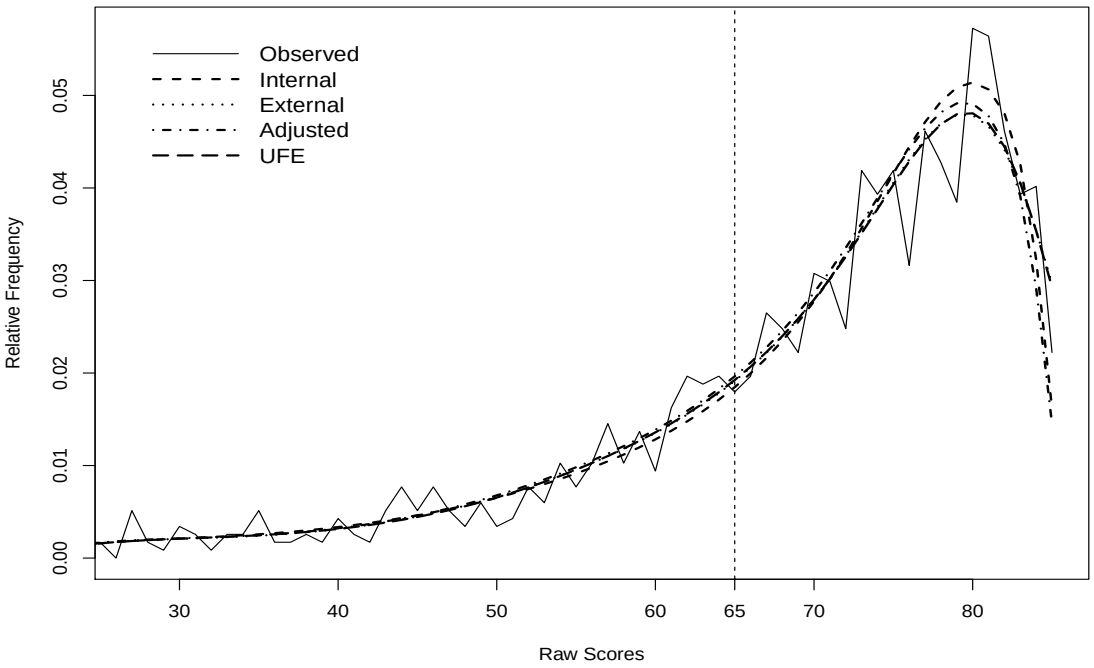


Figure 4.2: Test L1: Smoothed Relative Frequency Distributions

(a) Form X



(b) Form Y

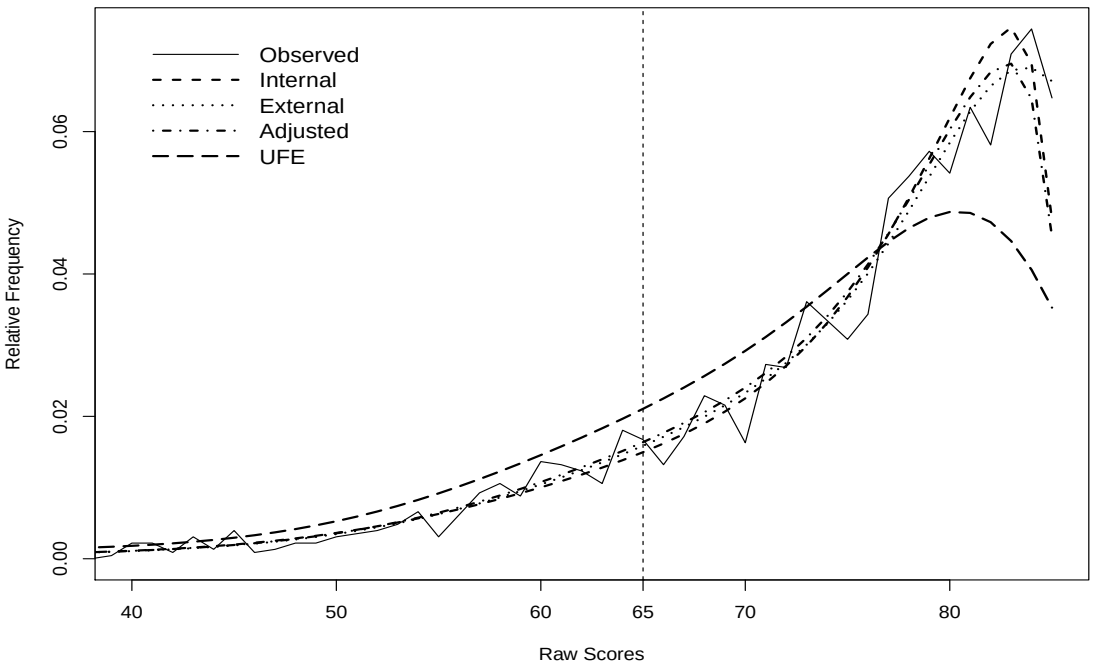
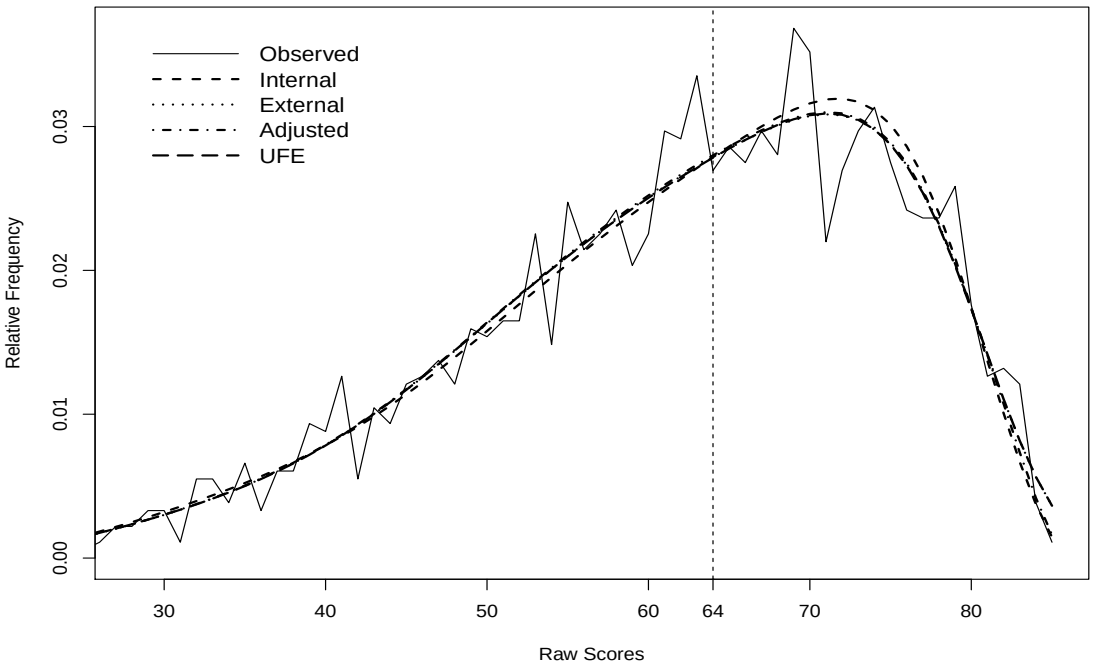


Figure 4.3: Test L2: Smoothed Relative Frequency Distributions

(a) Form X



(b) Form Y

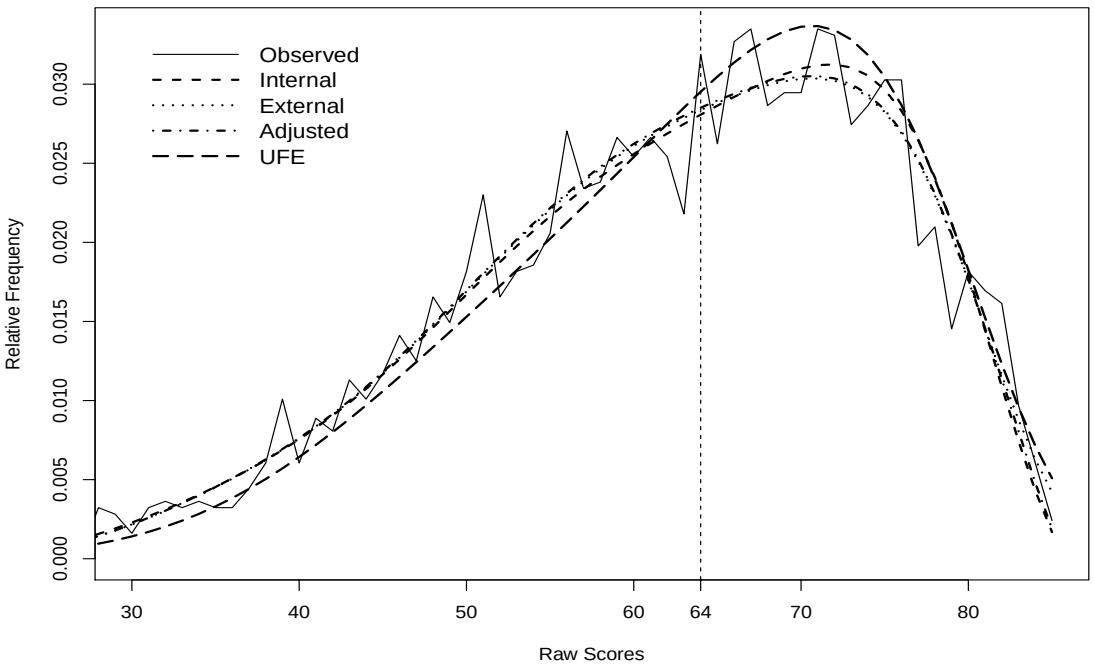
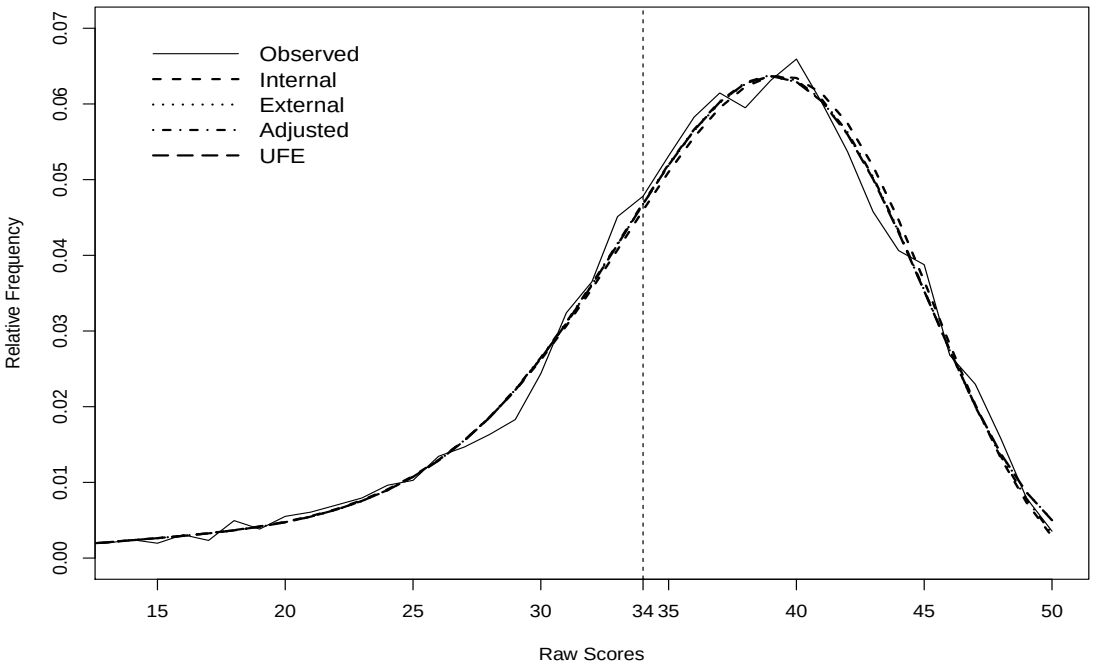


Figure 4.4: Test M: Smoothed Relative Frequency Distributions

(a) Form X



(b) Form Y

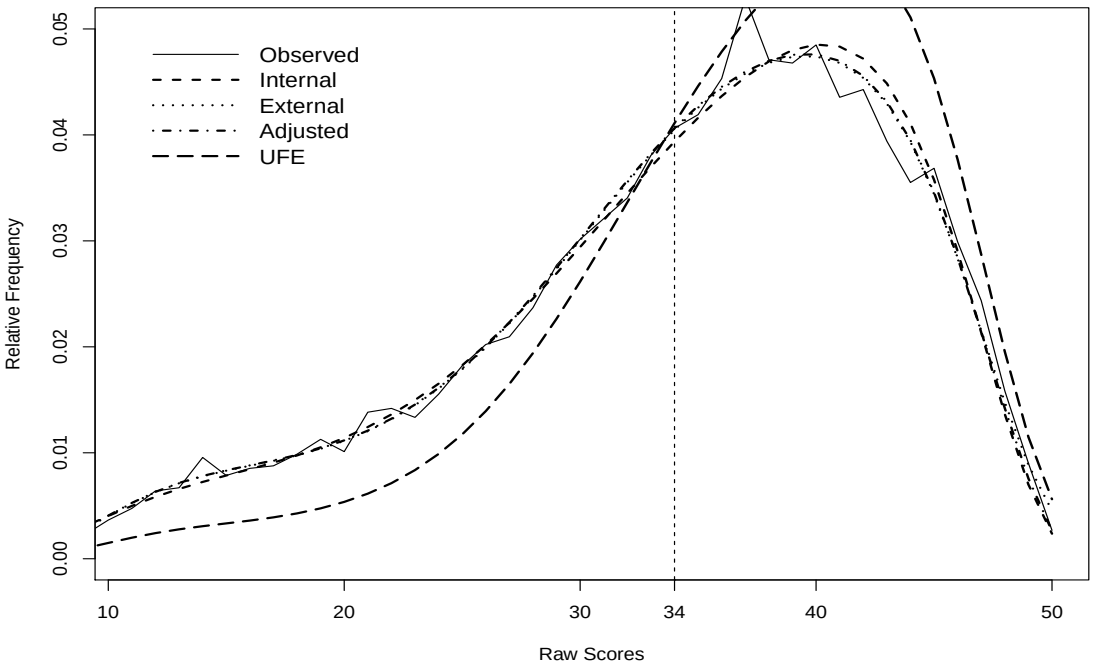
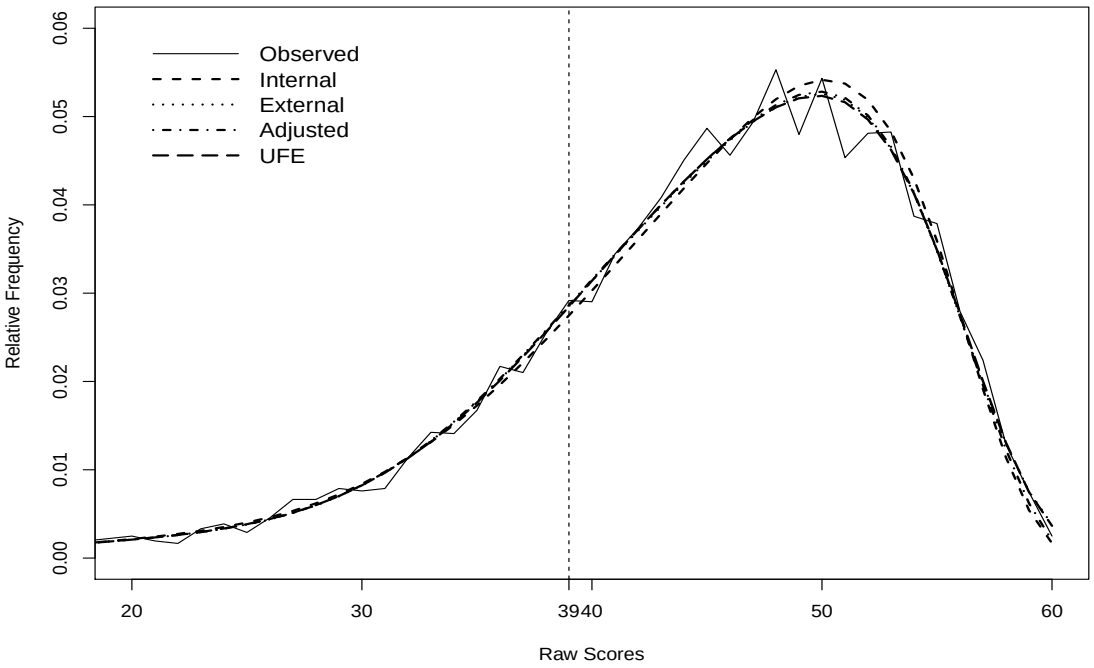


Figure 4.5: Test S1: Smoothed Relative Frequency Distributions

(a) Form X



(b) Form Y

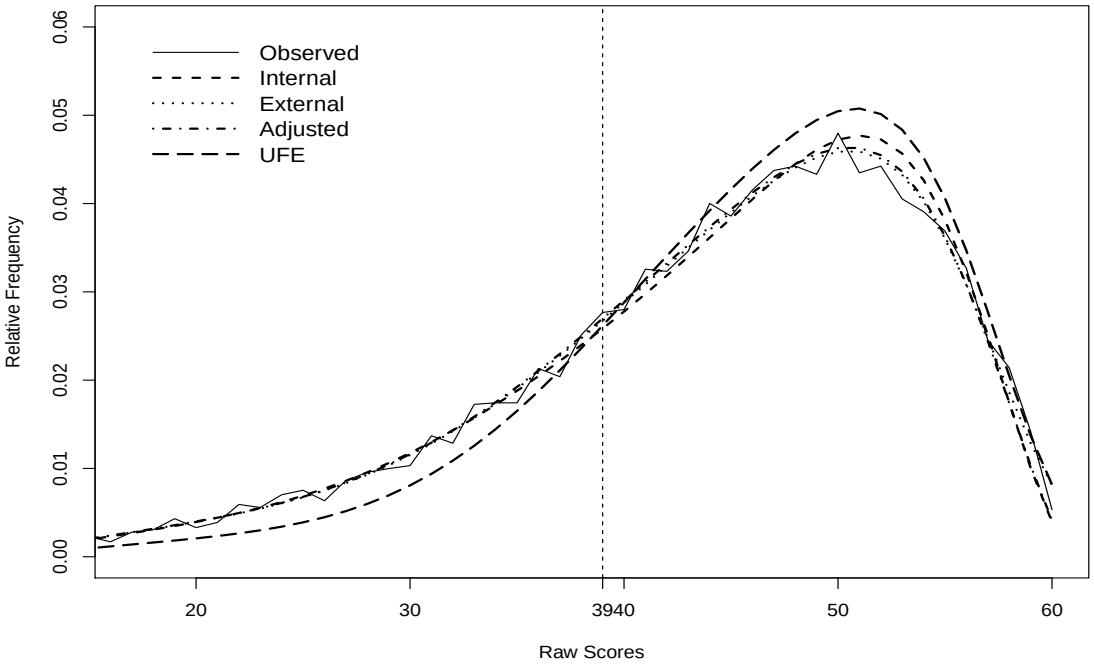
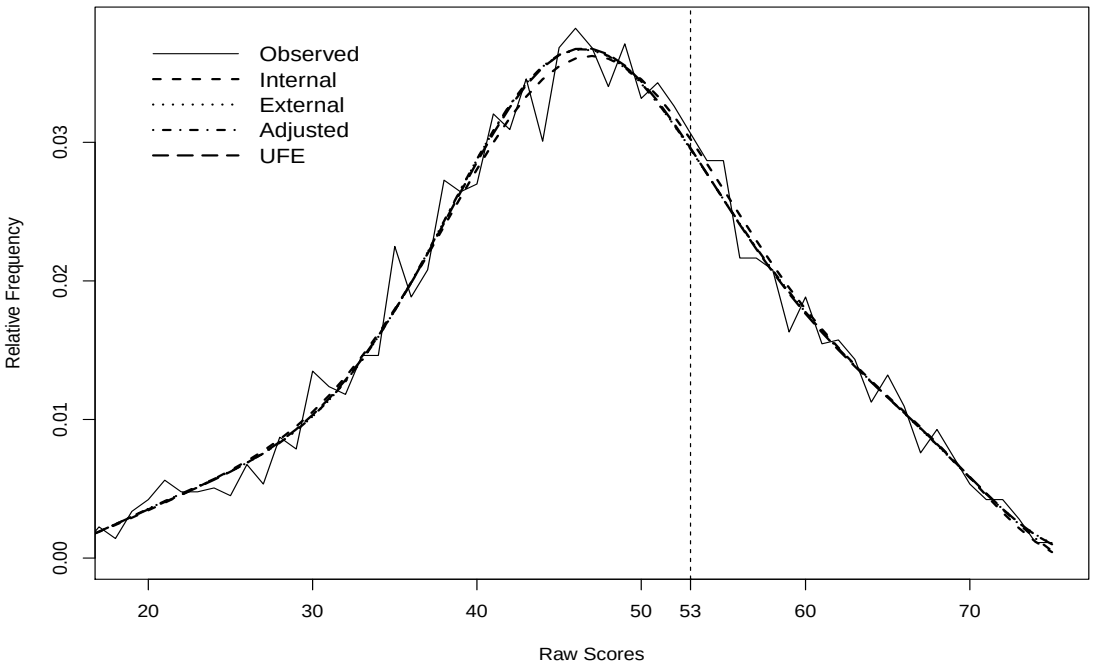


Figure 4.6: Test S2: Smoothed Relative Frequency Distributions

(a) Form X



(b) Form Y

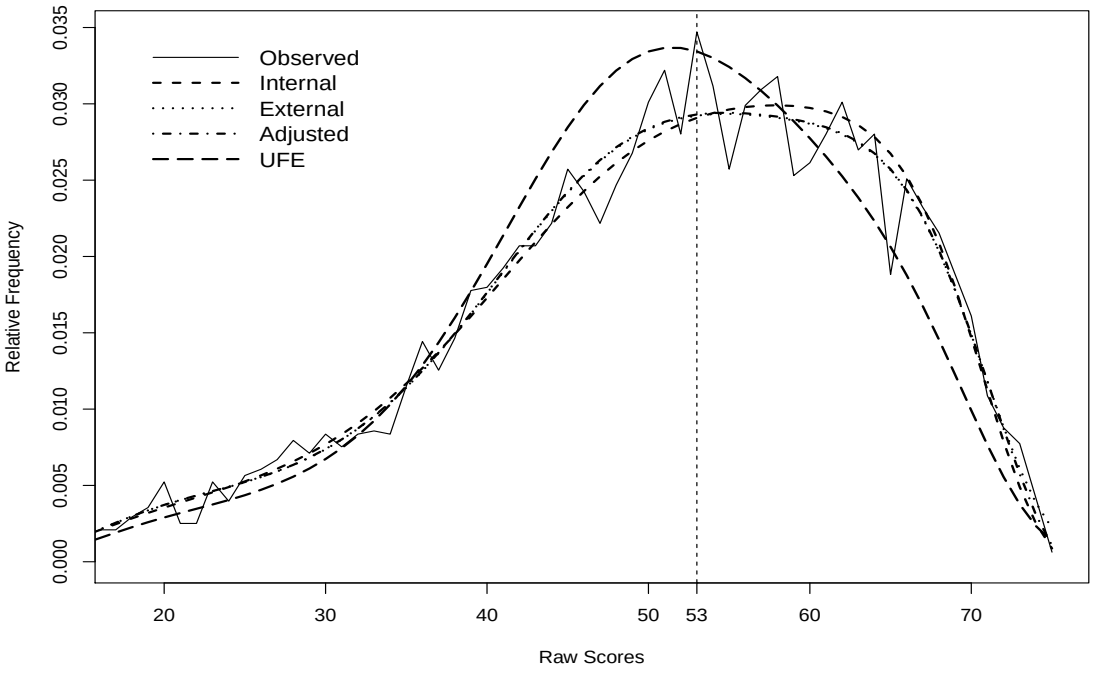
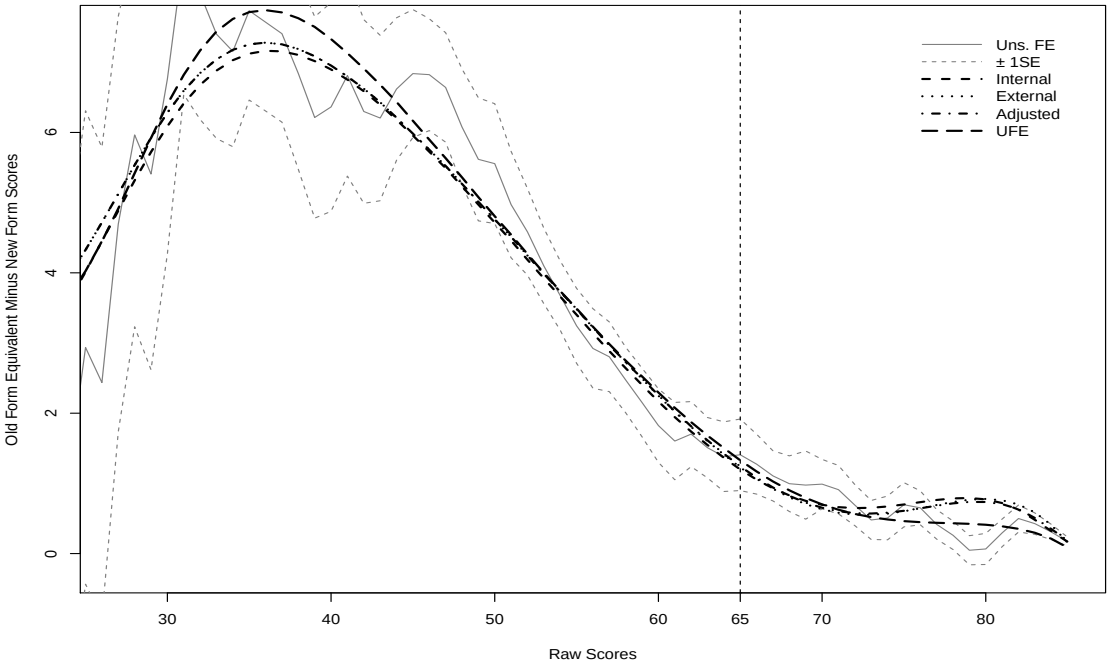


Figure 4.7: Test L1: Equating Relationship

(a) FE Method



(b) MFE Method

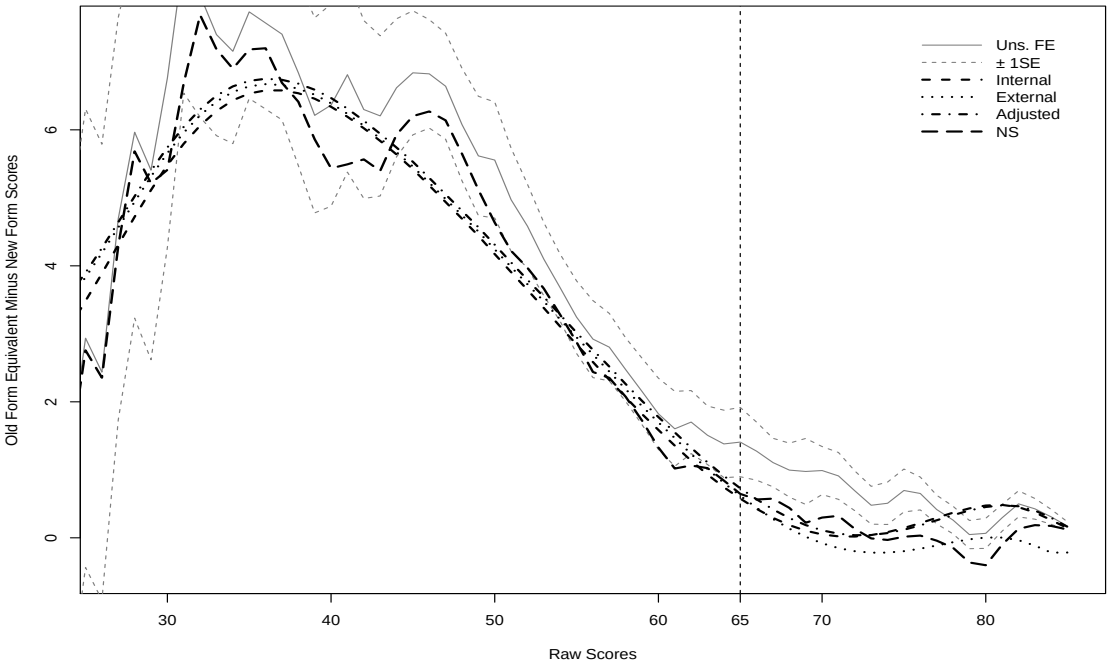
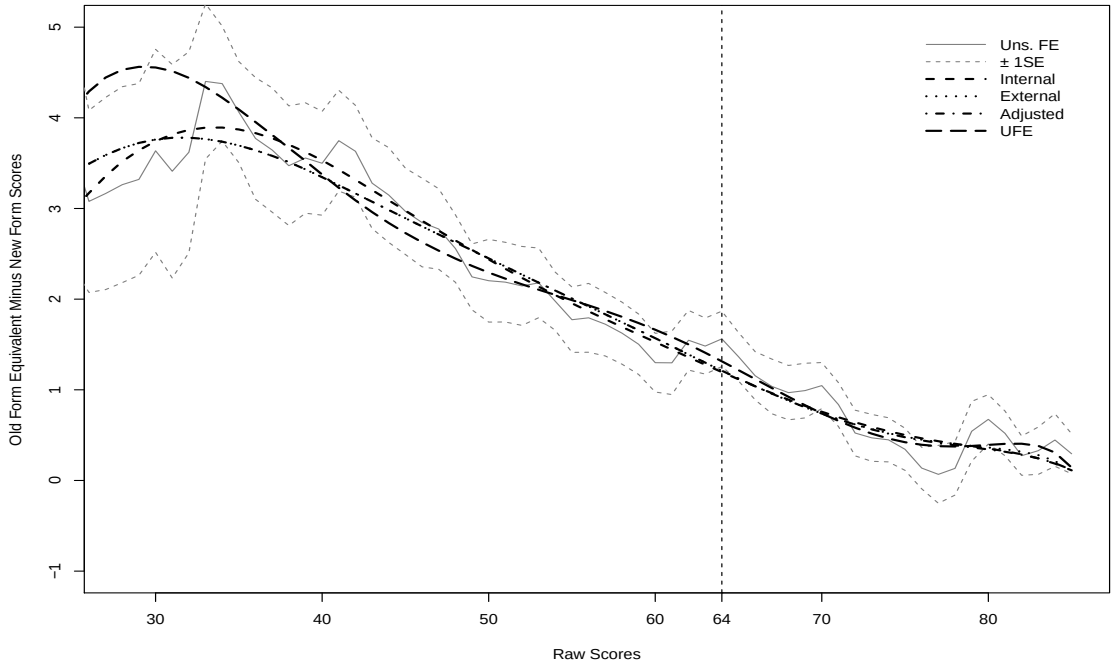


Figure 4.8: Test L2: Equating Relationship

(a) FE Method



(b) MFE Method

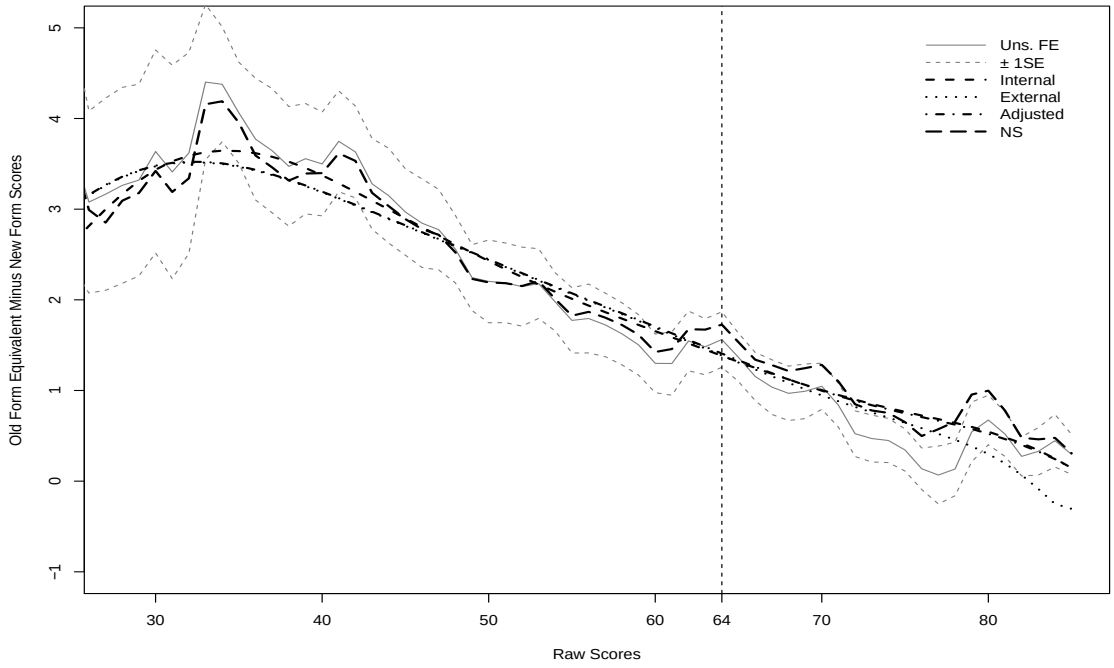
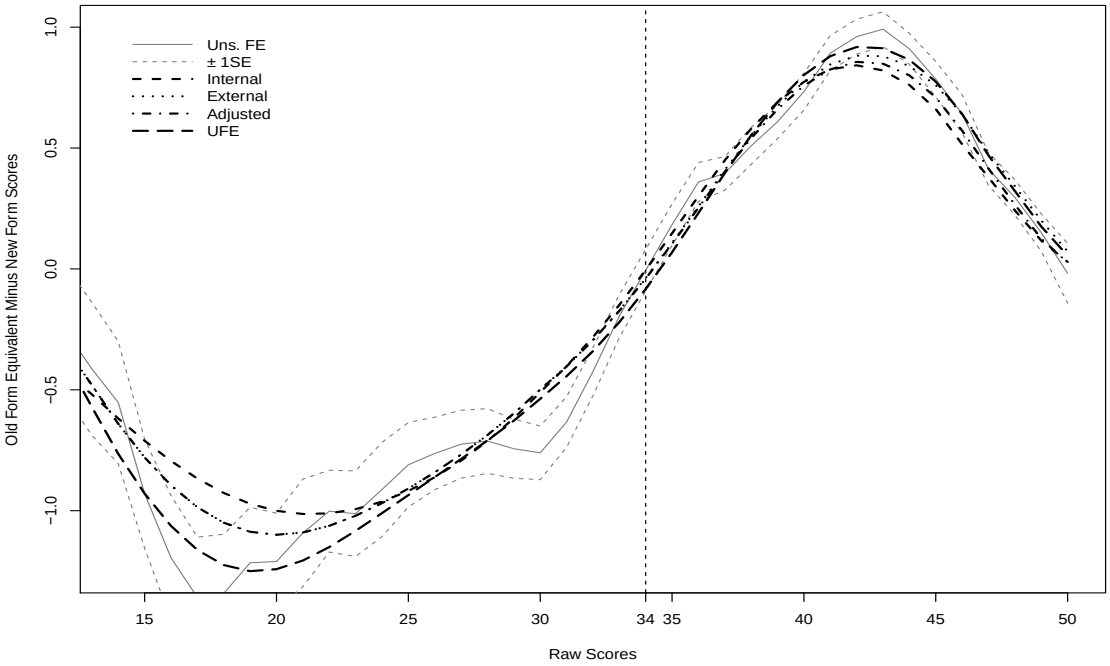


Figure 4.9: Test M: Equating Relationship

(a) FE Method



(b) MFE Method

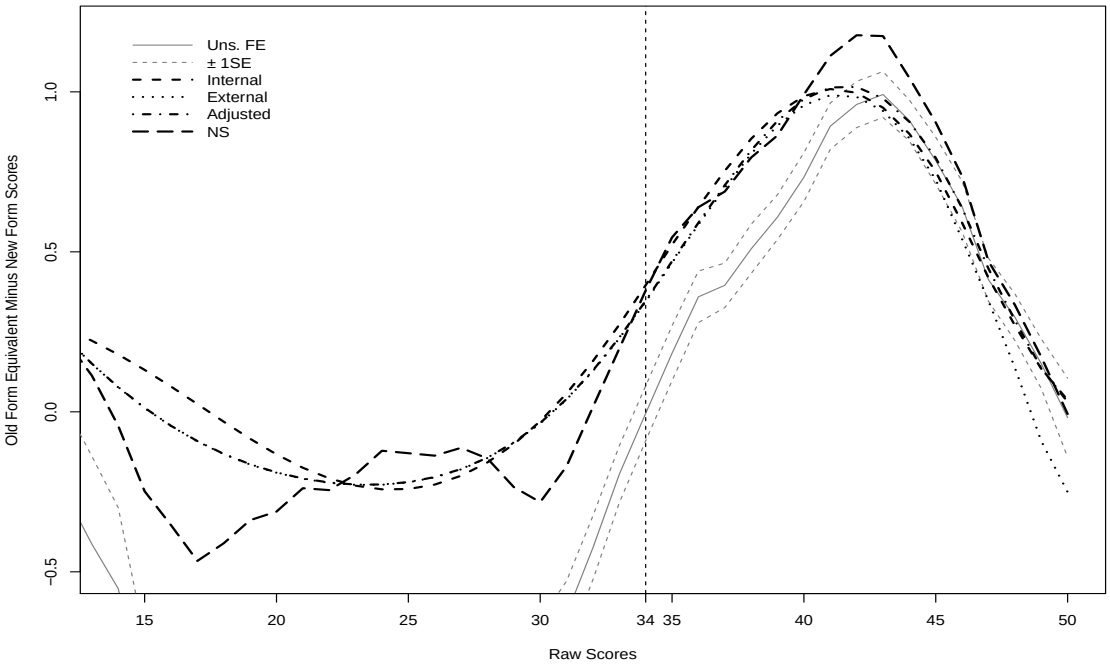
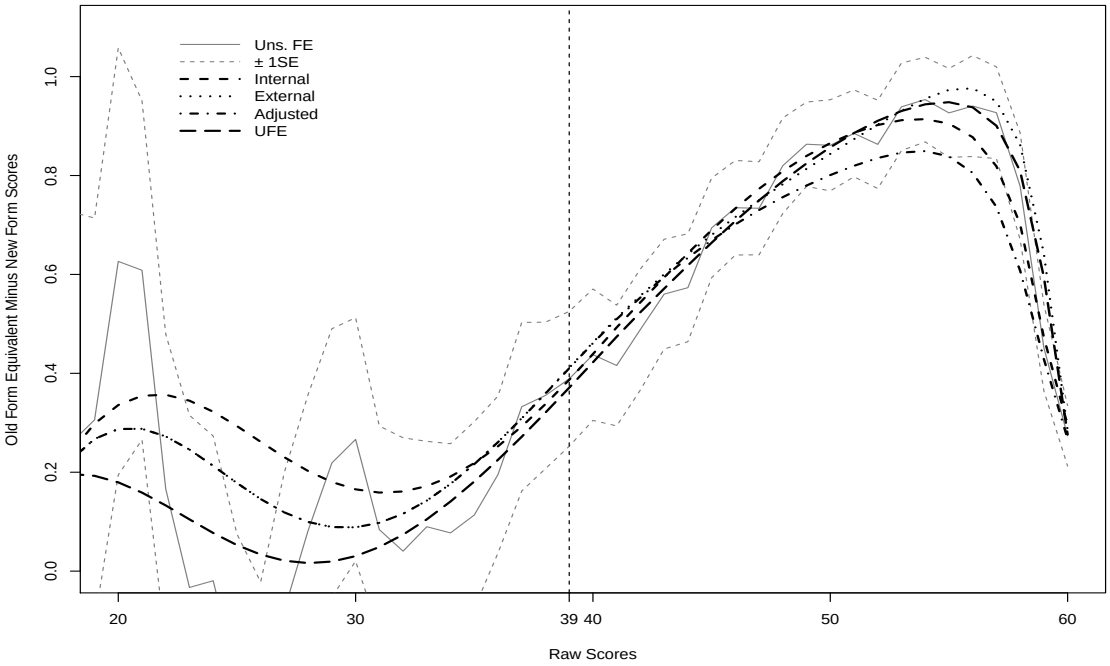


Figure 4.10: Test S1: Equating Relationship

(a) FE Method



(b) MFE Method

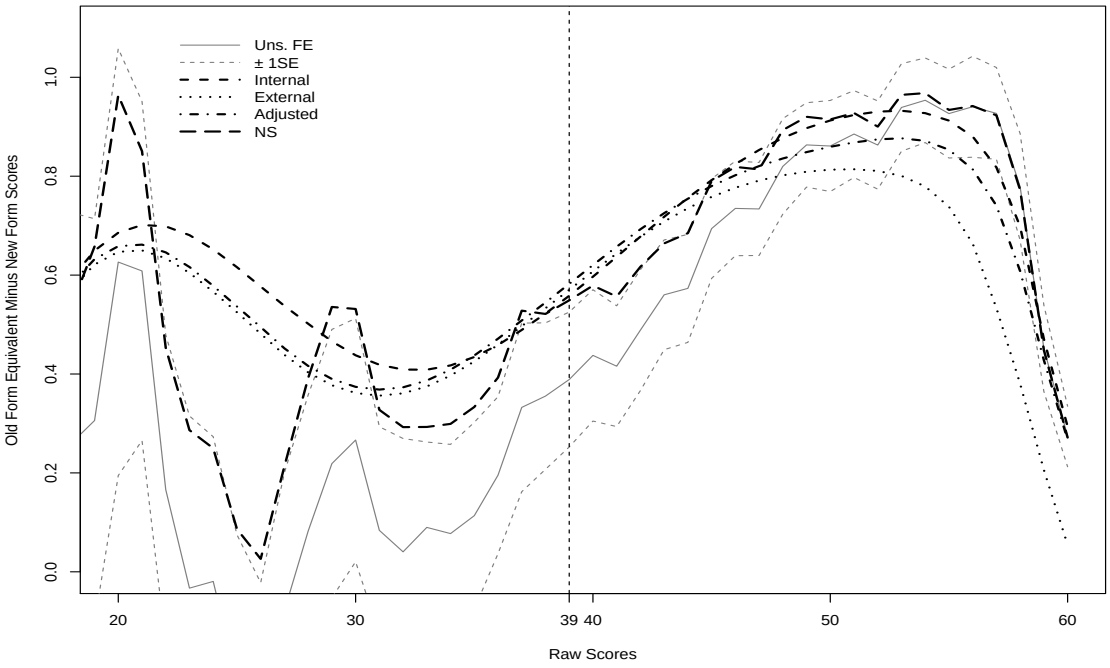
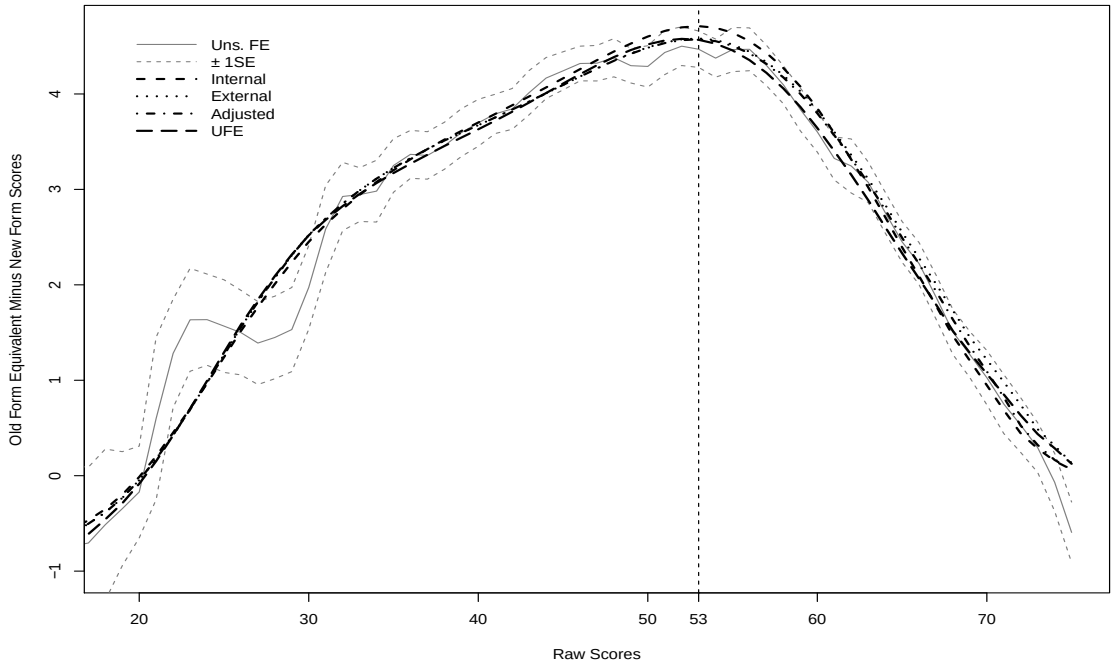


Figure 4.11: Test S2: Equating Relationship

(a) FE Method



(b) MFE Method

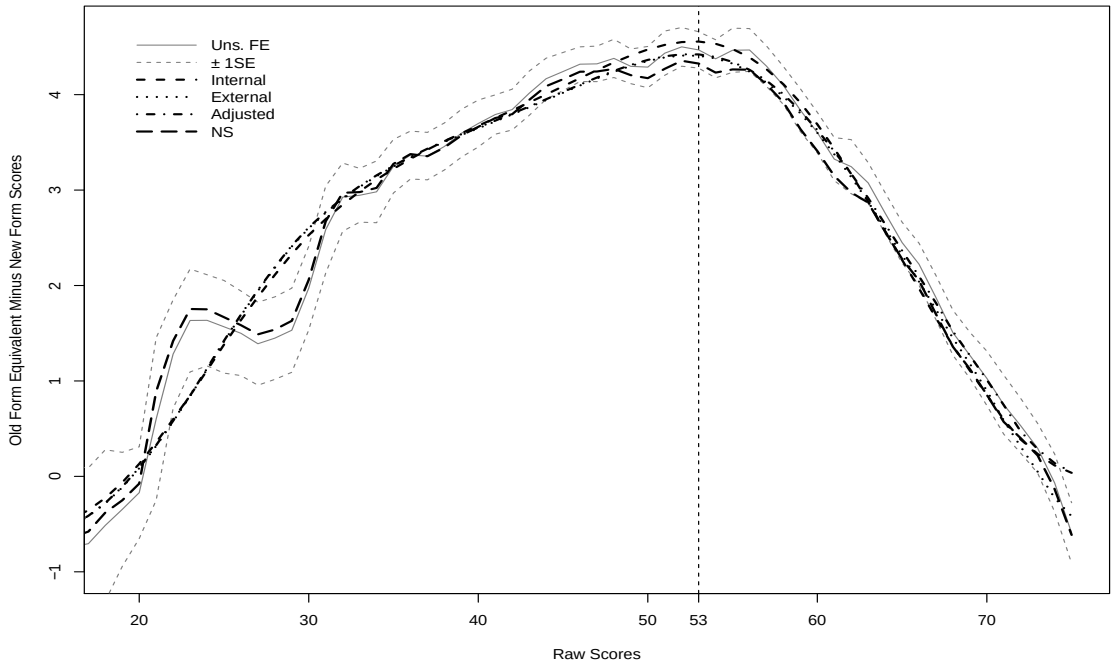
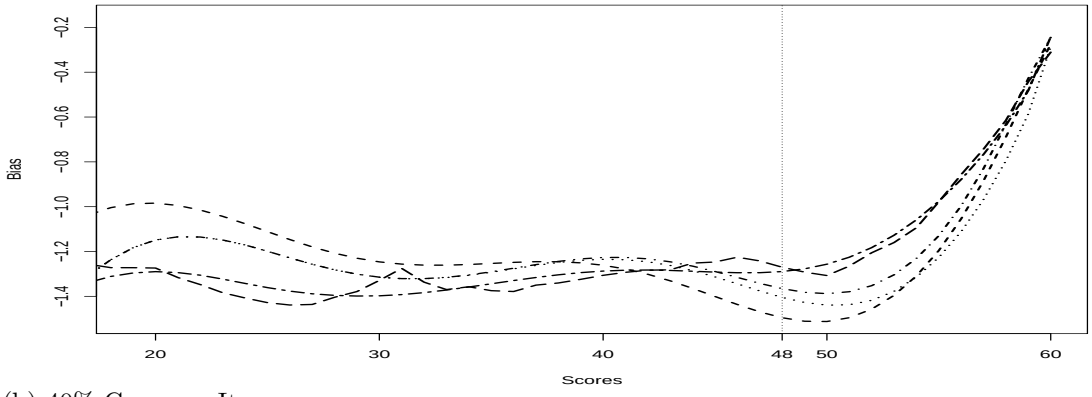
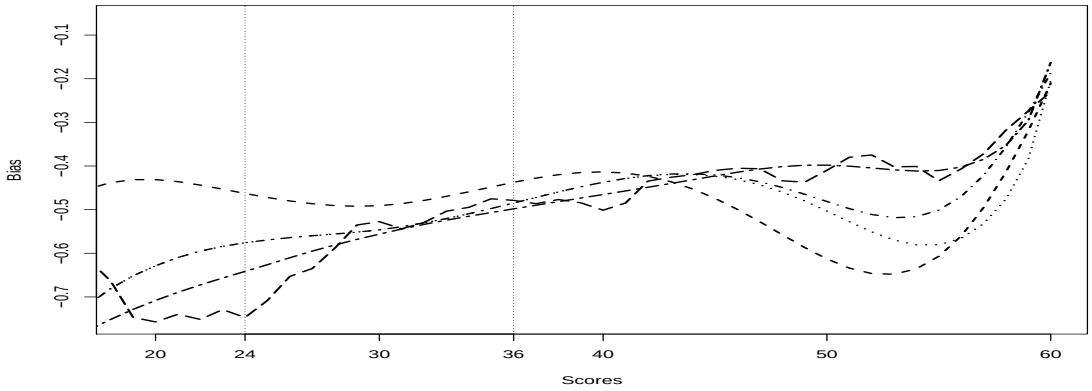


Figure 4.12: FE Method: Comparing Bias among Different Approaches (60 Items, Effect Size 0.50, Sample Size (1000, 1000))

(a) 20% Common Items



(b) 40% Common Items



(c) 60% Common Items

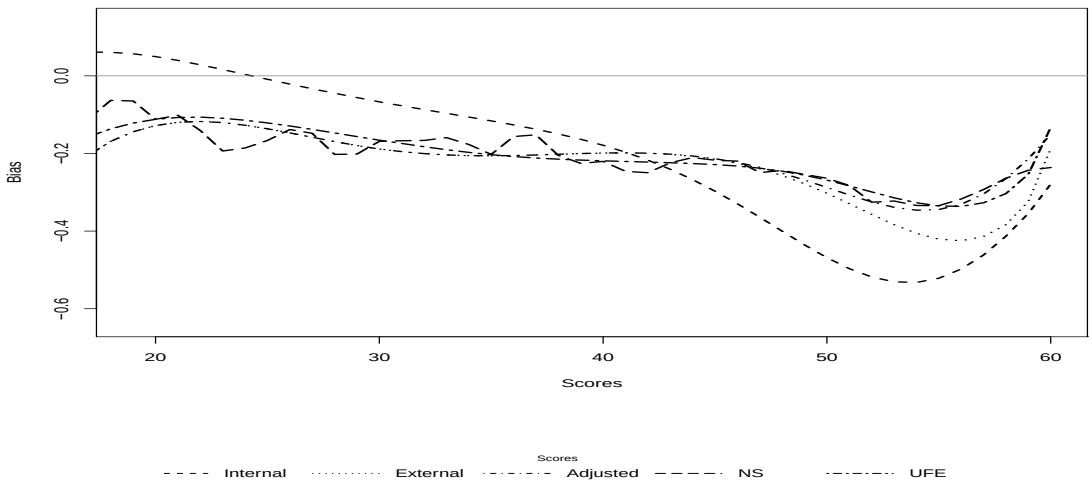
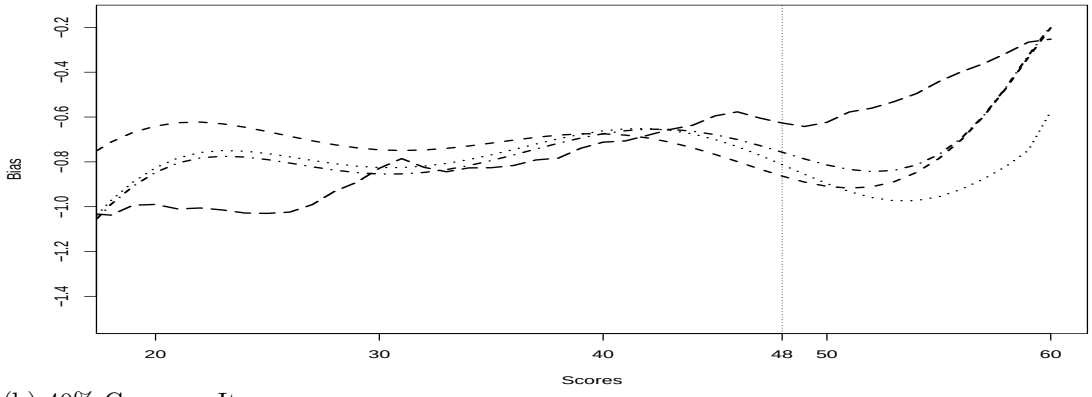
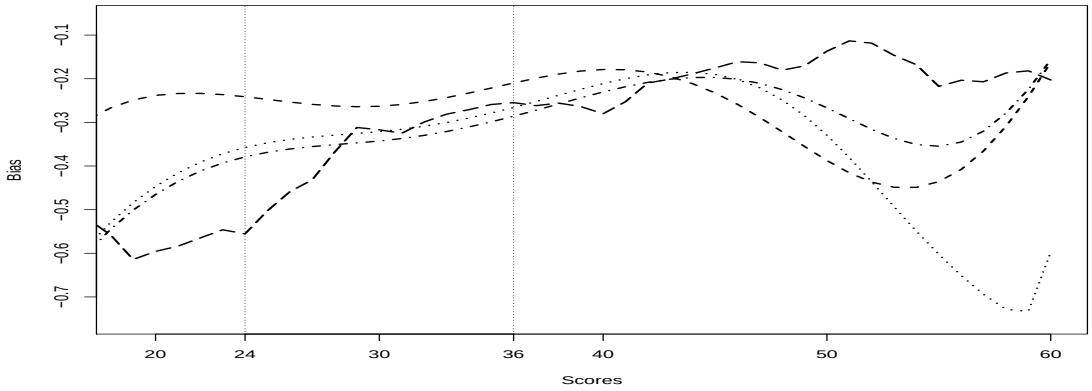


Figure 4.13: MFE Method: Comparing Bias among Different Approaches (60 Items, Effect Size 0.50, Sample Size (1000, 1000))

(a) 20% Common Items



(b) 40% Common Items



(c) 60% Common Items

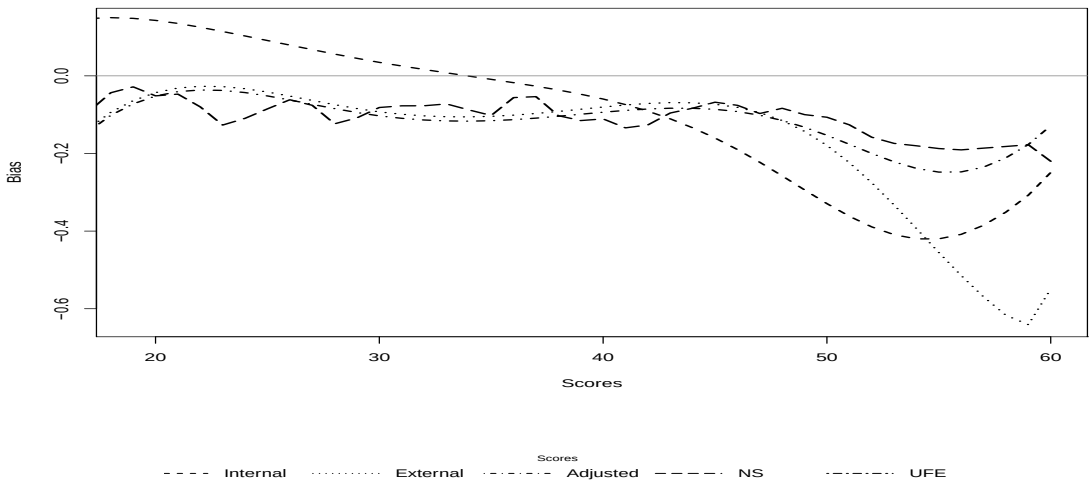
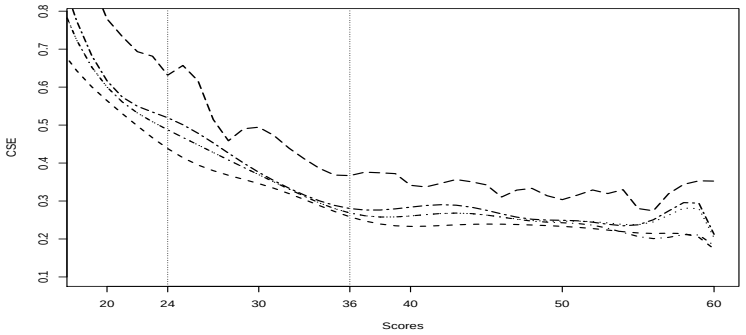


Figure 4.14: Conditional Statistics (CSE) Comparing Different Approaches to Handling Structural Zeros (60 Items, 40% Common Items, Effect Size 0.50, Sample Size (1000, 1000))

(a) FE Method



(b) MFE Method

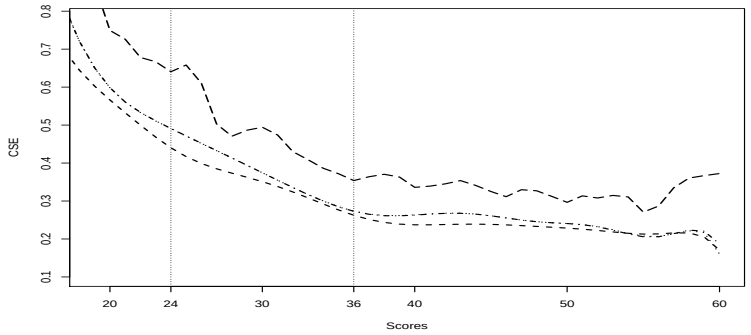
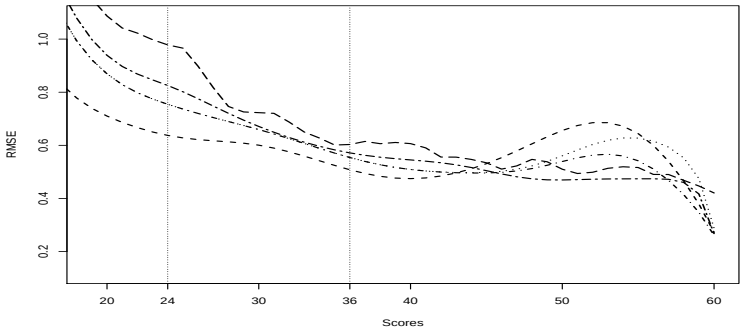
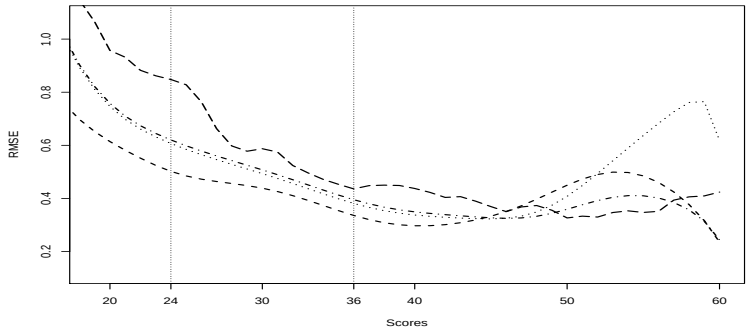


Figure 4.15: Conditional Statistics (RMSE) Comparing Different Approaches to Handling Structural Zeros (60 Items, 40% Common Items, Effect Size 0.50, Sample Size (1000, 1000))

(a) FE Method



(b) MFE Method



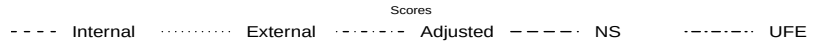


Figure 4.16: Effect of Proportion of Common Items (Unweighted Overall Statistics)

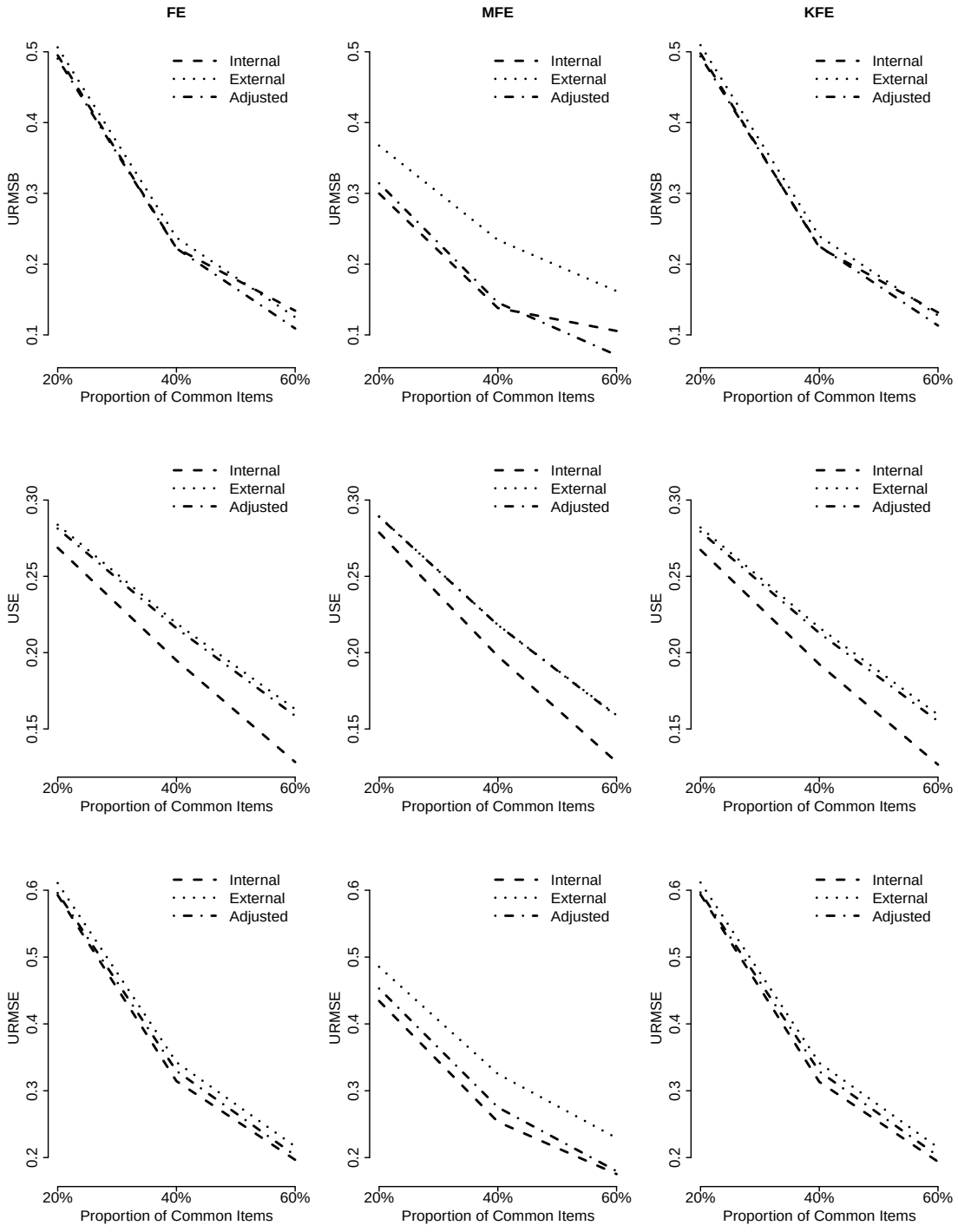


Figure 4.17: Effect of Proportion of Common Items (Weighted Overall Statistics)

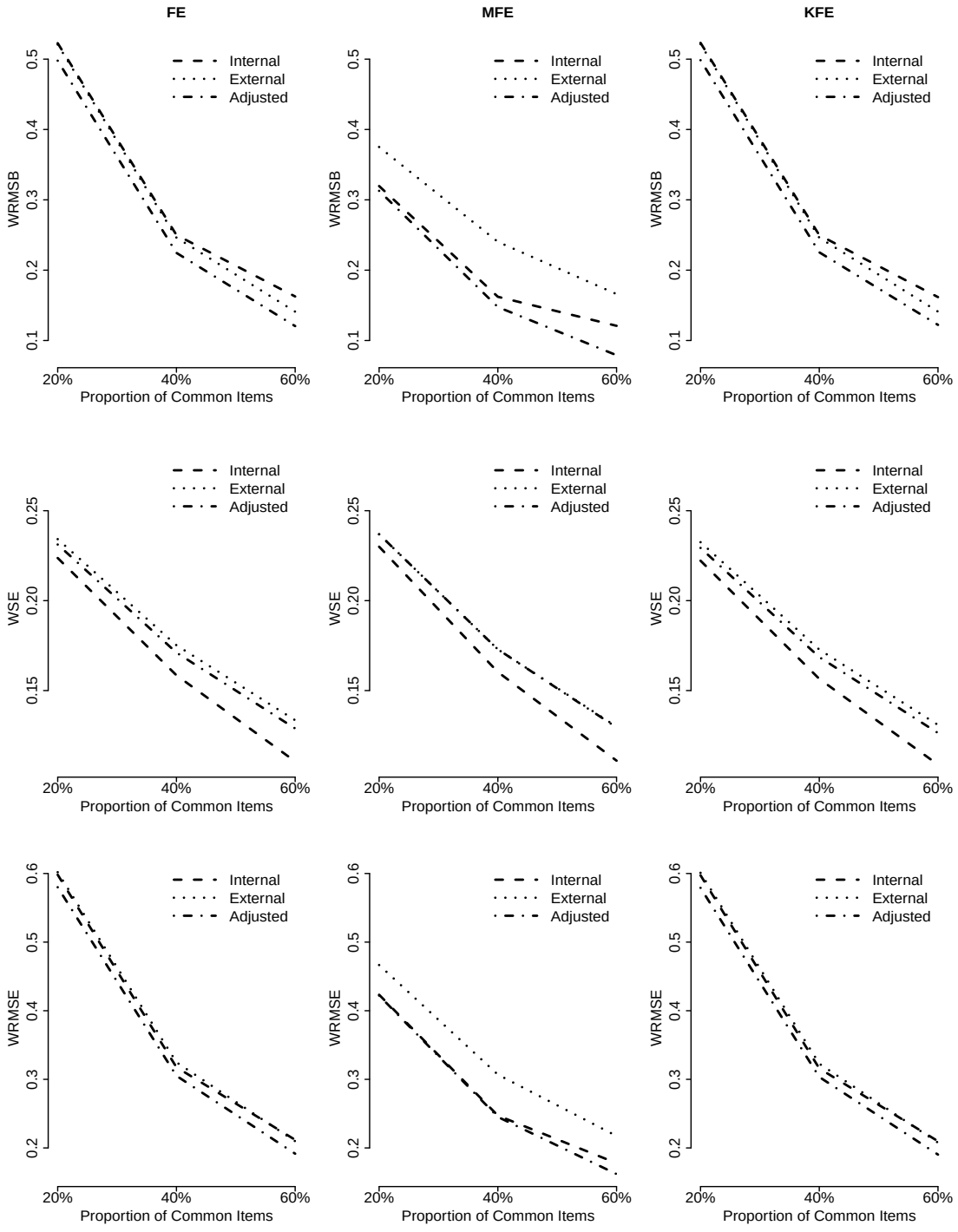


Figure 4.18: Effect of Test Length (Unweighted Overall Statistics)

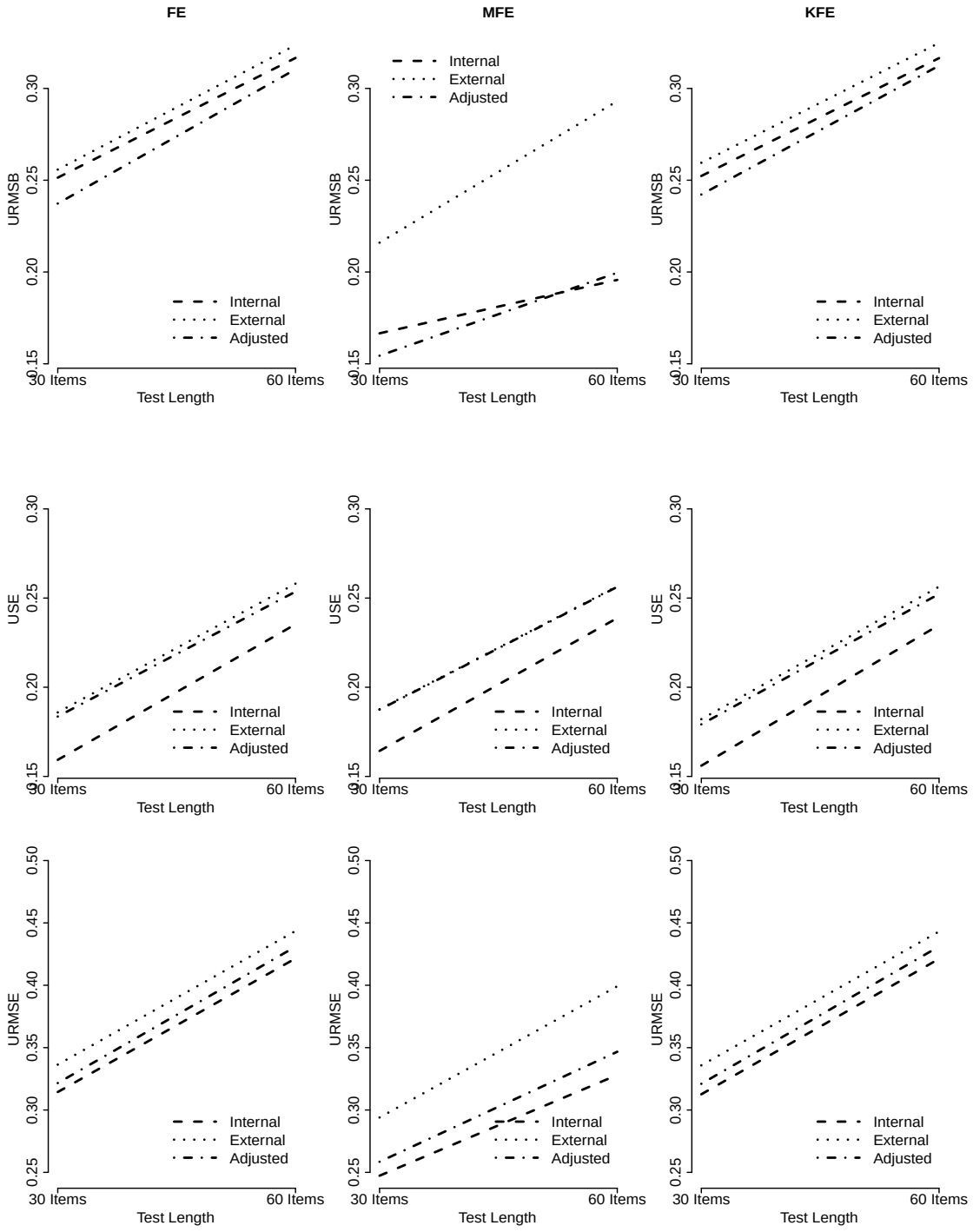


Figure 4.19: Effect of Test Length (Weighted Overall Statistics)

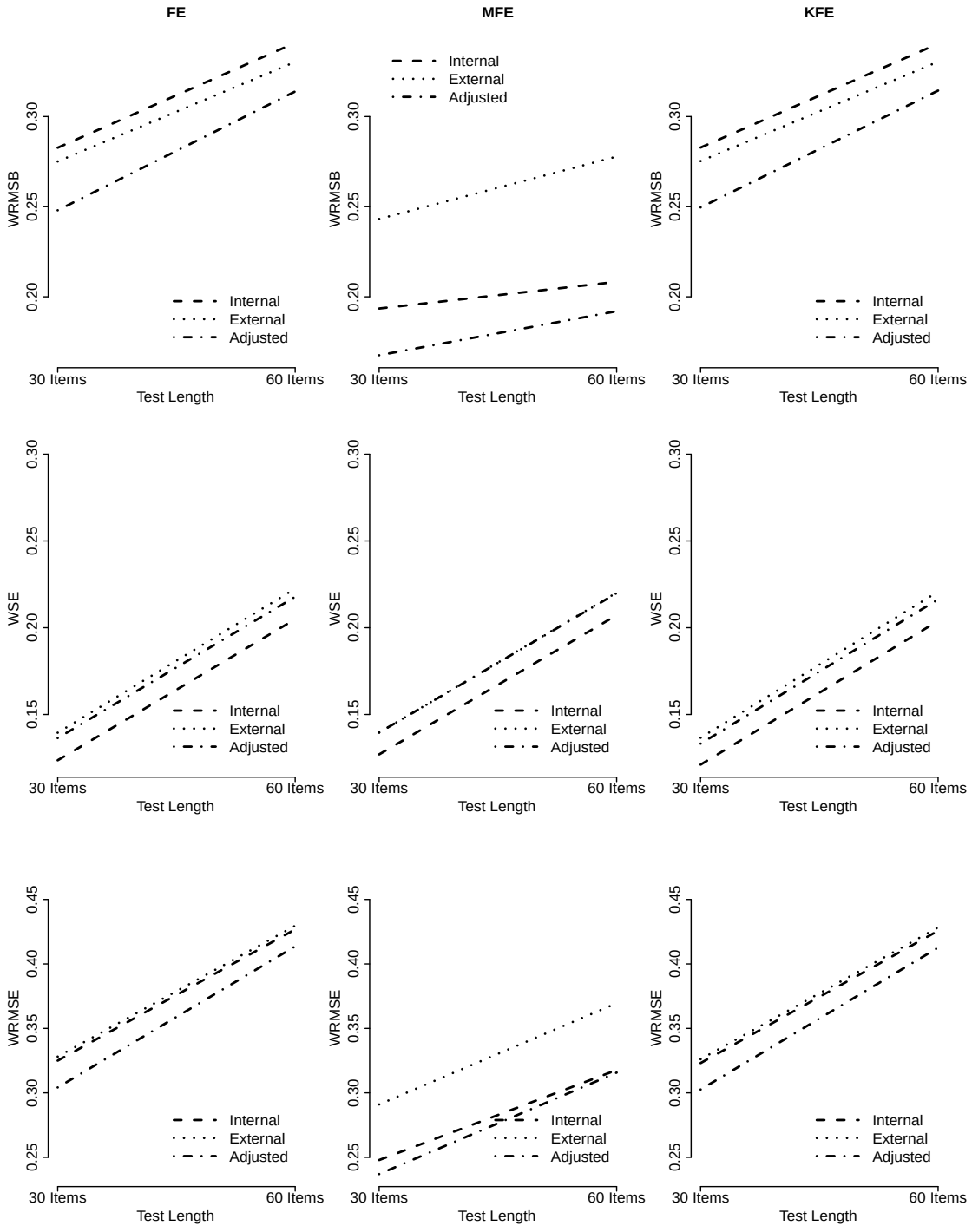


Figure 4.20: Effect of Effect Size (Unweighted Overall Statistics)

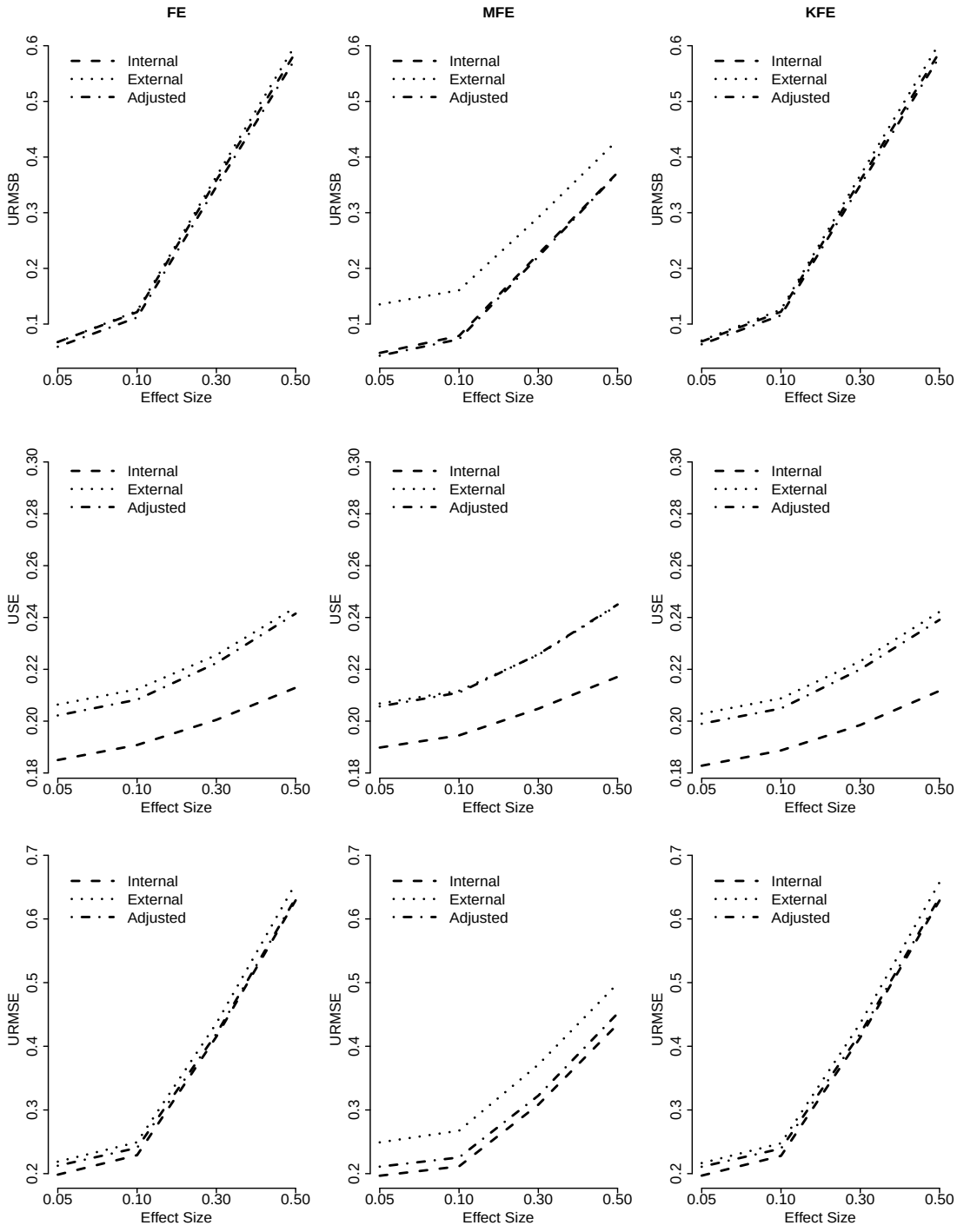


Figure 4.21: Effect of Effect Size (Weighted Overall Statistics)

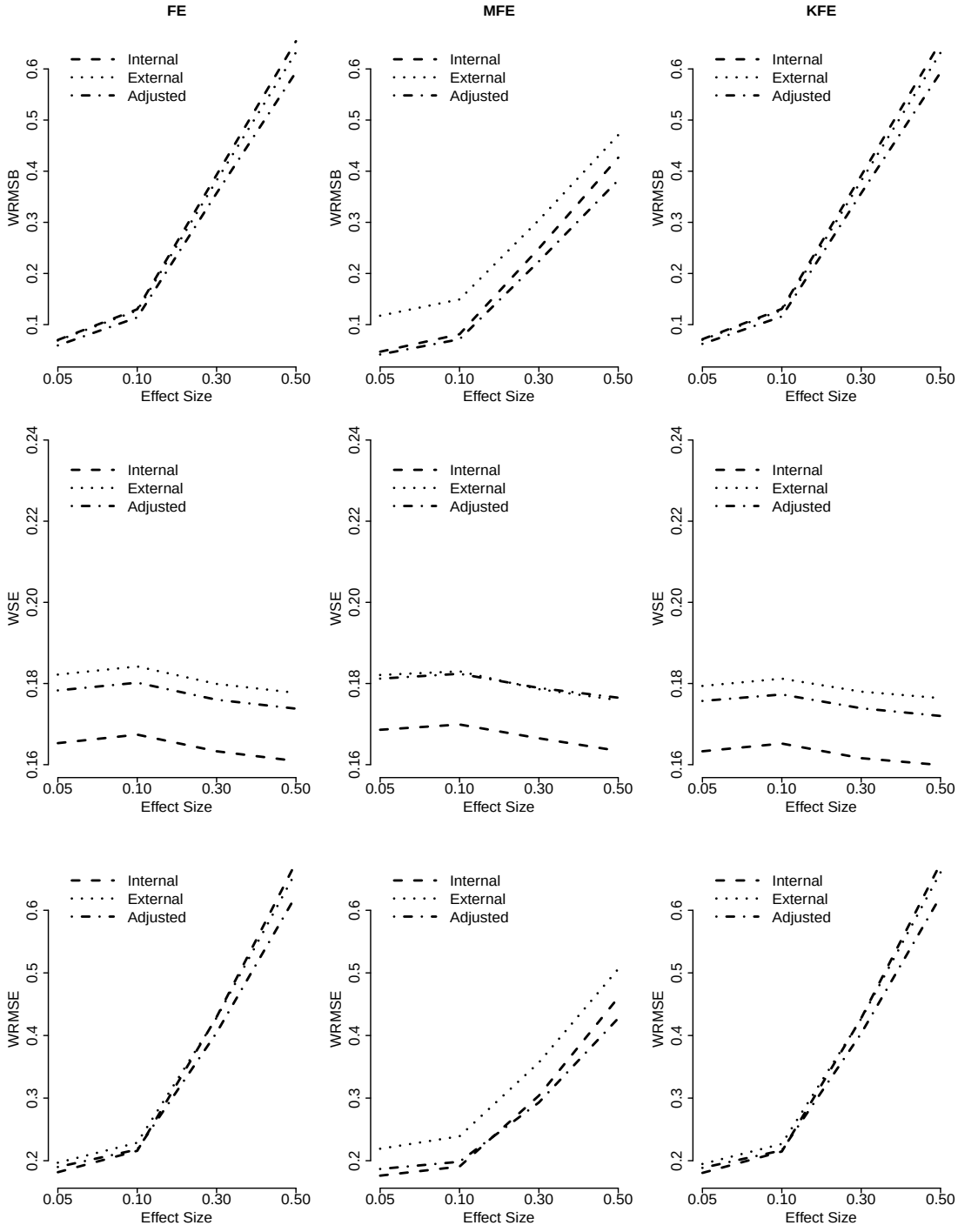


Figure 4.22: Effect of Sample Size (Unweighted Overall Statistics)

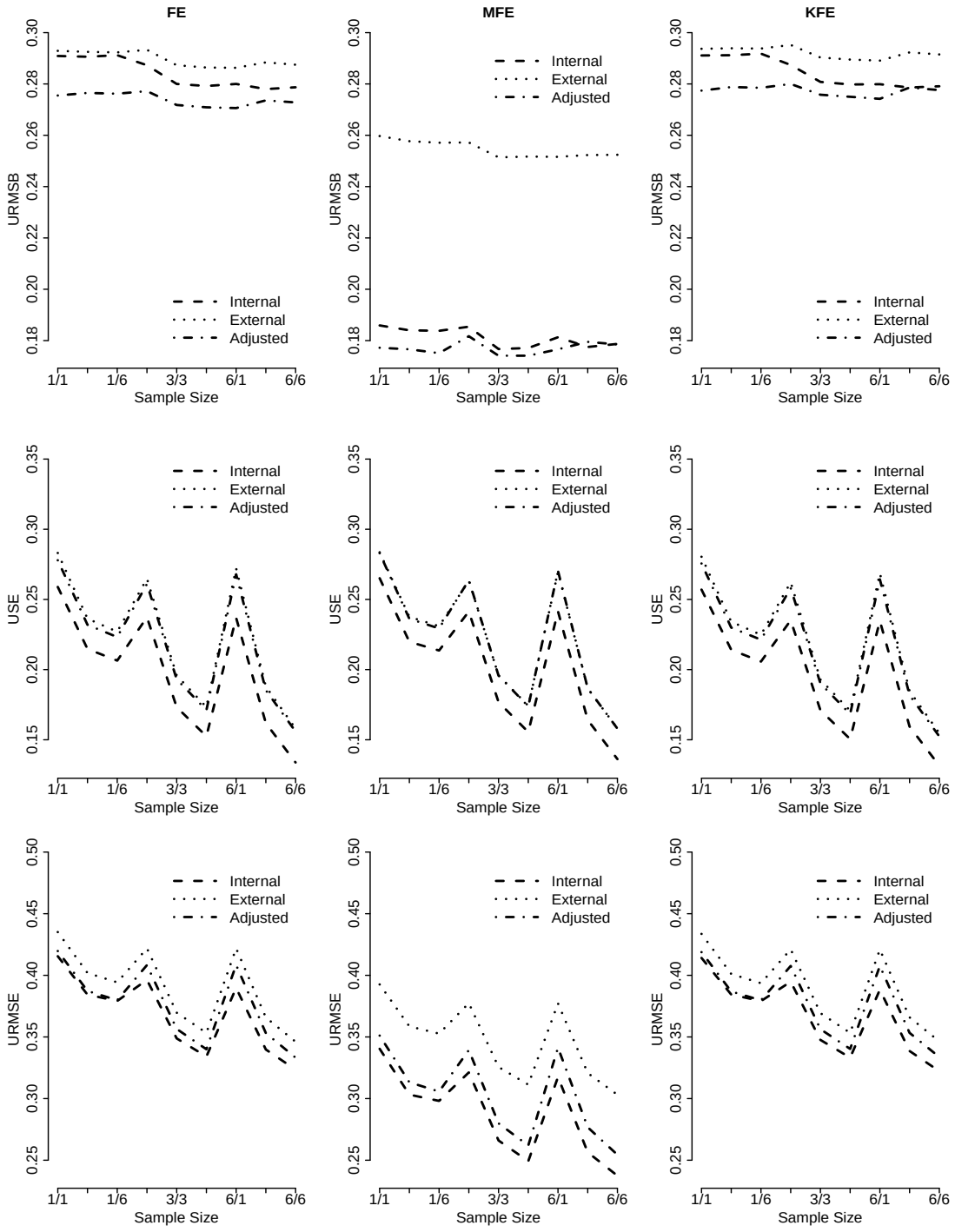
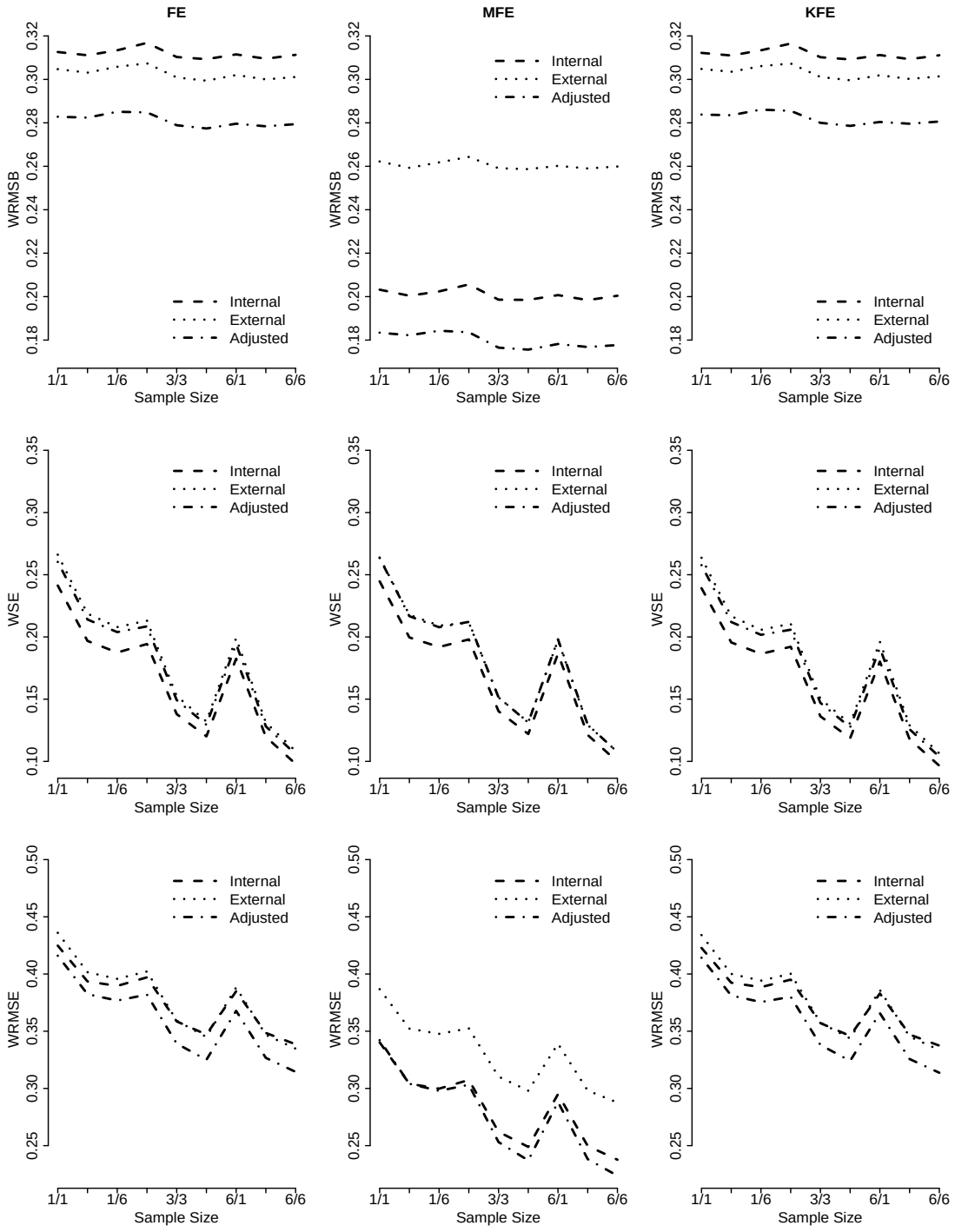


Figure 4.23: Effect of Sample Size (Weighted Overall Statistics)



References

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (2007). *Discrete multivariate analysis: Theory and practice*. New York, NY: Springer Science+Business Media, LLC.

Brennan, R. L., Wang, T., Kim, S., & Seol, J. (2009). *Equating Recipes* (Tech. Rep.). [Sofrware version 1.0]. Iowa City: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Hagge, S. L. (2010). *The impact of equating method and format representation of common items on the adequacy of mixed-format test equating using nonequivalent groups* (Unpublished doctoral dissertation). University of Iowa.

Holland, P. W., & Wang, Y. J. (1987). Regional dependence for continuous bivariate densities. *Communications in Statistics - Theory and Methods*, *16*(1), 193-206.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer Science+Business Media.

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

Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (2003). *BILOG-MG* (Tech. Rep.). [Computer Software]. Mooresville, IN: Scientific Software International.

APPENDIX A: Claim 1

Claim 1: The number of structural zeros is equal to $K_V(K_V+1)$.

Proof:

Situations where structural zeros occur can be divided into two different cases. Under each case, the number of structural zeros is found. Note that K_V, K_U, and K_X represent numbers of common items, non-common items, and total items, respectively, and that V and X represent scores on common items and total items, respectively.

Case 1) $X < V < K_V$

$$\begin{array}{lll} X=0 & V=1,2,\ldots,K_V & K_V \text{ counts} \\ X=1 & V=2,3,\ldots,K_V & K_V-1 \text{ counts} \\ \vdots & \vdots & \vdots \\ X=K_V-1 & V=K_V & 1 \text{ counts} \end{array}$$

Therefore, the number of structural zeros is

$$= \frac{K_V(K_V+1)}{2}.$$

Case 2) $K_U < X - V < K_X - V \Leftrightarrow K_U + V < X < K_X$

$$\begin{array}{llll} V=0 & X > K_U & \Rightarrow X = K_U+1,\ldots,K_X & K_X-K_U \text{ counts} \\ V=1 & X > K_U+1 & \Rightarrow X = K_U+2,\ldots,K_X & K_X-K_U-1 \text{ counts} \\ \vdots & \vdots & \vdots & \vdots \\ V=K_V-1 & X > K_U+K_V-1 = K_X-1 & \Rightarrow K_X & 1 \text{ counts} \\ V=K_V & X > K_U+K_V = K_X & & 0 \text{ counts} \end{array}$$

Therefore, the number of structural zeros is

$$= \frac{(K_X-K_U)(K_V+1)}{2} = \frac{K_V(K_V+1)}{2}$$

because $K_X = K_V + K_U$.

Combining the numbers of structural zeros from both Case 1 and Case 2 makes the total number of structural zeros equal to

$$K_V(K_V+1).$$

APPENDIX B: Claim 2

Claim 2: The proportion of structural zeros can be approximately estimated by the proportion of common items.

Proof:

First of all, the total number of possible pairs of the total score and the common item score based on the $X \times V$ matrix is $(K_X + 1)(K_V + 1)$. Since it is known that the number of structural zeros equal to $K_V(K_V + 1)$ from Claim 1, the proportion of structural zeros becomes

$$\begin{aligned} \frac{K_V(K_V+1)}{(K_X+1)(K_V+1)} &= \frac{K_V}{(K_X+1)} \\ &\approx \frac{K_V}{K_X} \\ &= \text{the proportion of common items.} \end{aligned}$$

Therefore, the proportion of structural zeros can be approximately estimated by the proportion of common items. And, the proportion of structural zeros becomes closer to the proportion of common items as the number of total items (K_X) increases.

APPENDIX C: Claim 3

Claim 3: When the internal approach is used to smooth bivariate distributions, the first and second moments (i.e., mean and standard deviation) for smoothed distributions are the same as the observed mean and standard deviation as long as the bivariate degree of smoothing is at least $(2, 2, 1, 1)$.

Proof:

Suppose that the internal approach is applied on the $U \times V$ score distribution where $X = U + V$. For the observed data, let

$\hat{\mu}_x =$ actual mean for scores on X
$\hat{\mu}_v =$ actual mean for scores on V
$\hat{\mu}_u =$ actual mean for scores on U
$\hat{\sigma}_x =$ actual standard deviation for scores on X
$\hat{\sigma}_v =$ actual standard deviation for scores on V
$\hat{\sigma}_u =$ actual standard deviation for scores on U
$\hat{\sigma}_{xv} =$ actual covariance between scores on X and V
$\hat{\sigma}_{xu} =$ actual covariance between scores on X and U

For the internal approach, let

${}_i\mu_x =$ mean for scores on X
${}_i\mu_v =$ mean for scores on V
${}_i\mu_u =$ mean for scores on U
${}_i\sigma_x =$ standard deviation for scores on X
${}_i\sigma_v =$ standard deviation for scores on V
${}_i\sigma_u =$ standard deviation for scores on U
${}_i\sigma_{xv} =$ covariance between scores on X and V
${}_i\sigma_{xu} =$ covariance between scores on X and U.

Then, by the definition of the internal approach,

$$ {}_i\mu_u = \hat{\mu}_u,\ {}_i\mu_v = \hat{\mu}_v,\ {}_i\sigma_u = \hat{\sigma}_u,\ {}_i\sigma_v = \hat{\sigma}_v, \text{ and } {}_i\sigma_{xu} = \hat{\sigma}_{xu}, $$

Let's consider ${}_i\mu_x$, ${}_i\sigma_x$, and ${}_i\sigma_{xv}$.

$$ {}_i\mu_x = \mathrm{E}(U + V) = \mathrm{E}(U) + \mathrm{E}(V) = {}_i\mu_u + {}_i\mu_v = \hat{\mu}_u + \hat{\mu}_v = \hat{\mu}_x. $$

Therefore, ${}_i\mu_x = \hat{\mu}_x$.

$$ \begin{aligned} {}_i\sigma_x &= \mathrm{Var}(X) = \mathrm{Var}(U + V) \\ &= \mathrm{Var}(U) + \mathrm{Var}(V) = 2\mathrm{Cov}(U, V) \\ &= {}_i\sigma_u^2 + {}_i\sigma_v^2 + {}_i\sigma_{uv} \\ &= \hat{\sigma}_u^2 + \hat{\sigma}_v^2 + \hat{\sigma}_{uv} = \hat{\sigma}_x^2. \end{aligned} $$

Thus, ${}_i\sigma_x^2 = \hat{\sigma}_x^2$. Therefore, for the internal approach, the first and second moments are preserved the same as the observed moments.