Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 47

Subscore Equating and Reporting

Euijin Lim
Won–Chan Lee[†]

May 2016

[†]Euijin Lim is a research assistant in the Center for Advanced Studies in Measurement and Assessment (CASMA), College of Education, University of Iowa (email: euijin-lim@uiowa.edu). Won-Chan Lee is Director, Center for Advanced Studies in Measurement and Assessment (CASMA), 210 Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (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

The purpose of this study is to address the necessity of subscore equating and to evaluate the performance of various equating methods for subtests. Assuming the random groups design and number-correct scoring, this research analyzed real data and simulated data with four study factors including test dimensionality, subtest length, form difference in difficulty, and sample size. Various traditional and IRT equating methods were considered.

The main findings of this study are as follows: (1) reporting subscores without equating provides misleading information in terms of score profiles; (2) reporting subscores without a pre-specified test specification brings practical issues such as constructing alternate subtest forms with comparable difficulty, conducting equating between forms with different lengths, and deciding an appropriate score scale to be reported; (3) the best performing subscore equating method, overall, was IRT observed score equating using 3PL with separate calibration (3PsepO) followed by equipercentile equating with presmoothing, and the worst performing method was IRT true score equating using BF with simultaneous calibration (BFT); (4) simultaneous calibration involving other subtest items in the calibration process yielded larger bias but smaller random error than did separate calibration, indicating that borrowing information from other subtests increases bias but decreases random error in subscore equating; (5) IRT observed score equating using BF with simultaneous calibration (BFO) performs the best when a test is multidimensional, form difference is small, subtest length is short, or sample size is small; (6) equating results for BFT and BFO were affected by the magnitude of factor loading and variability for the estimated general and specific factors; and (7) smoothing in general improved equating results.

1 Introduction

When a test consists of several different content areas or constructs, scores on the subdomains are called subscores. Usually a test blueprint does not consider a situation where subscores are reported; however, there is a non-negligible demand for subscore reporting in the testing industry. Policy makers, admission officers, educators, and individual task takers want to receive subscores to make decisions about remediation or classification (Brennan, 2011; Haberman, 2008). Due to such demands, many testing programs already include subscores as a part of their score report or consider a plan of reporting subscores. For example, the College Board (2014) announced that, for their redesigned SAT, they are planning to report seven subscores to provide more insight into students strengths and weaknesses. ACT (2014) also included three to ten subscores called category scores for each subject in the ACT Aspire score report.

Ideally, a decision on subscore reporting should be made at the very early stage of test development; unfortunately, subscores are often added after a testing program has been defined and used. In such cases, a test specification does not include detailed requirements for subtests, so subtests would have different numbers of items between subtests and even between forms. In order to facilitate a comparison of scores from different content areas, scores are typically expressed in a score profile—a common unit that has a norm-referenced meaning (Thorndike & Thorndike-Christ, 2009). Even for tests with subscores on different scale score metrics, bar charts or line graphs are used for the subscores in order to enable a comparison of performances across subareas. For a test not originally designed to report subscores, it is common to report proportion-correct subscores or percent subscores without equating.

However, using a profile or putting subscores on a proportion-correct score scale is not enough to make a valid comparison of subscores. Once the decision to report subscores has been made, equating is a necessary process to sustain the meaning of subscores across forms given that a test is commonly administered with multiple alternate forms. For subscores, there is a more serious issue related to equating: a score profile based on unequated subscores could inform an examinee's relative performance on different subtests in reverse. That is, if subscores are not equated, relative strengths and weaknesses of examinees shown in score profiles do not solely show their proficiencies, but reflect differences in subtest form difficulties as well. For example, a student's score profile for mathematics could say that he is better at algebra than at geometry when using unequated scores; however, his geometry score could become higher than his algebra score after equating.

Subtests have two salient properties: short length and relatedness with other subtests belonging to the same test. Constructing alternate forms of comparable difficulty is challenging for subscores that typically have a limited number of items because form difficulty is substantially affected by the choice of each item (Stahl & Masters, 2009). Consequently, equating subscores can be challenging, too. The other property of subtests is that subscores are correlated with each other. Because a subtest consists of items that are a subset of the total test,

subscores are correlated with the other subscores and the total scores. The relatedness property makes it possible to use information from items in other subtests to stabilize scoring or equating for a target subtest.

There are only a few existing studies in the literature that deal with subscore equating. Sinharay and Haberman (2011) suggested equating methods for regressed subscores, in which a target subtest was scored by borrowing information from the other subtests. Even though number-correct scoring is still the most widely used in operational setting, a thorough study dealing with subscore equating for number-correct scores has not been conducted. Moreover, the effect of equating on score profiles has not been studied yet.

For this research, the random groups (RG) design is used and ten equating methods, including traditional and IRT equating, are compared with each other and with predefined criteria. This research consists of two studies as to the purposes and data types. The first study involves applying various equating methods to operational datasets to show how score profiles change, particularly when compared with unequated results. The main purpose of the first study is to address the necessity of subscore equating. The second study compares and evaluates the performance of various equating methods for subscores using simulated data.

2 IRT Equating for Subscores

Kolen and Brennan (2014) describe equating using IRT as a three-step process: item parameter calibration, scale transformation, and equating for number-correct scores. Scale transformation is usually not required under the RG design because item parameters calibrated using equivalent groups are already on the same scale. For equating using MIRT, there is another issue related to scale linking: the scales may vary due to rotational indeterminacy. Brossman and Lee (2013) and E. Lee (2013) showed that, under the RG design, results for multidimensional IRT (MIRT) observed score equating are invariant to rotation since the fitted marginal observed score distributions are not affected by rotation.

In terms of the calibration step, there are two possible approaches in IRT subscore equating. One is separate calibration for each subtest, which treats a subtest as an independent test and rules out a possible adverse effect of other subtests. The other approach is simultaneous calibration for the entire test, which in a sense entails borrowing information from other subtests. For simultaneous calibration, it is important to consider whether to use a unidimensional IRT (UIRT) or MIRT model. If a test consists of multiple subtests but they are very similar to each other, using a UIRT model should provide acceptable results. For a test with very distinctive subtests, on the other hand, applying a MIRT model seems reasonable.

IRT observed score equating is equipercentile equating using IRT model-fitted score distributions for the old and new forms. A conditional distribution of observed scores given a certain ability level θ_i is estimated using a recursive

formula (Lord & Wingersky, 1984). For Bi-factor (BF) observed score equating (G. Lee & Lee, 2016), marginal distributions are computed using a MIRT version of the Lord-Wingersky formula.

IRT true score equating uses test characteristic curves (TCC) for the old and new forms to equate number-correct scores. The challenge of BF true score equating is to construct a unidimensional TCC. Under MIRT, each item has an item characteristic surface (ICS) instead of an item characteristic curve (ICC), and the sum of ICSs for items belonging to a test is a test characteristic surface (TCS). While a TCC relates a single value of θ to a single true score, a TCS maps various combinations of θs to a single true score. That is, multiple θ may correspond to the same true score. G. Lee et al. (2015) suggest a BF true score equating method that constructs unidimensional TCCs by integrating out specific dimensions. For each item, this method first integrates the ICS over the specific dimension. Then, each item has an ICC conditional on the general ability parameter only. The sum of the ICCs across a test becomes a TCC. Once TCCs for the old and new forms are constructed, UIRT true score equating can be applied.

3 Method

3.1 Study 1: Operational Data Analyses

This study uses data from large-scale French Language (FL) and Physics (PHY) tests, which both have multiple-choice and constructed-response items. Equating data were originally collected using the common-item nonequivalent groups (CINEG) design with an internal anchor; this study uses only the multiple-choice items in the tests. The tests are not designed to report subscores, but each test has multiple content domains. The multiple-choice section of FL is divided into two domains: reading and listening. The first 30 items belong to the reading section, and 7 of them are common items (CI). The other 35 items belong to the listening section, and 15 of them are common. Some of the FL items share common reading or listening stimuli; in this research, testlet effect was not taken into account. The multiple-choice section of PHY consists of six content areas including classical mechanics, electricity and magnetism, waves and optics, thermal physics, fluid mechanics, and atomic and nuclear physics. The number of items in each content area varies from 3 to 25 and even differs between the old and new forms. The items belonging to each content area are interspersed throughout the test. The total number of multiple-choice items is 70 and the number of CIs is 21.

In order to sample pseudo-equivalent groups, this study considered matching techniques using background variables related to the performance on the tests. Background variables including gender, ethnicity, and parental education levels were considered. After recoding the background variables into 0/1 binary variables, exact matching and propensity score matching were conducted using all possible combinations of background variables. An *R* package *MatchIt* (Ho,

Imai, King, & Stuart, 2011) was used for matching. Each matched sample was evaluated in terms of the common-item effect size. For PHY, matching techniques did not appear to improve the equivalence between samples in terms of the effect size. For FL, exact matching using gender, ethnicity, and parental education slightly improved group equivalence. Following matching, a random sample of 3,000 examinees was repeatedly drawn for each form with a constraint of a small effect size to approximate the RG design. Pseudo-equivalent groups for FL were sampled considering not only the effect size of the whole test, but also the effect size of each domain; for PHY, in contrast, samples were drawn without consideration of the content areas because content areas with a small number of items have very few or no CIs between the old and new forms. Characteristics of the FL and PHY tests are provided in Table 1.

For this research, the following equating methods were considered: 1) identity equating, 2) linear equating, 3) equipercentile equating, 4) equipercentile equating with log-linear presmoothing, 5) equipercentile equating with cubic-spline postsmoothing, 6) IRT true score equating using 3PL with separate calibration (3PsepT), 7) IRT observed score equating using 3PL with separate calibration (3PsepO), 8) IRT true score equating using 3PL with simultaneous calibration (3PsimT), 9) IRT observed score equating using 3PL with simultaneous calibration (3PsimO), 10) IRT true score equating using BF with simultaneous calibration (BFT), and 11) IRT observed score equating using BF with simultaneous calibration (BFO).

The presmoothing and postsmoothing parameters considered for the subtests are shown in Table 2. For log-linear presmoothing, the smoothing parameter C was selected based on the difference chi-square statistic. For the fifth subtest of PHY which is the shortest one, the degree of presmoothing is the same as the subtest length for both forms, so the fitted distributions perfectly recover the observed score distributions; thus, the presmoothing equating result is the same as that of equipercentile equating without smoothing. For cubic-spline postsmoothing, equating results using various S values were compared and one of them was selected based on multiple criteria including ± 1 standard error bands, smoothness of a fitted equating function, moment differences between old-form scores and equated scores, and characteristics of scale scores. All the graphical and numerical comparisons of postsmoothing results were conducted using *ESUM-RG* (Brennan et al., 2015).

Equating Recipes (Brennan, Wang, Kim, & Seol, 2009) was used to conduct traditional equating. For IRT equating methods, *flexMIRT* (Cai, 2013) was used to estimate item parameters. After item parameters were estimated either simultaneously or separately, *Equating Recipes* was used to conduct all the IRT equating procedures except the BF equating methods. The BF equating methods were implemented using *R* (R Core Team, 2014).

The equating results of the operational data analyses were compared at two levels: the subtest level and the test level. For both levels, identity equating results were used as a baseline for comparison. Identity equating (i.e., no equating) is widely used in practice for subscore reporting even though it does not improve score comparability. Large differences between equated and unequated

results would suggest the necessity of subscore equating. At the subtest level, the weighted average of the absolute value of the differences (MAD) (Kolen & Brennan, 2014) between each equating method and the identity equating results were computed as:

$$MAD = \Sigma_j w_j |\hat{e}_Y(x_j) - x_j|, \tag{1}$$

where x_j is the unequated raw-score point, $\hat{e}_Y(x_j)$ is the equated score of x_j, and w_j is the relative frequency of x_j in the new-form group. Scores falling below the sum of c-parameter estimates were not included in the computation. Because the PHY subtests have different numbers of items between the old and new forms, x_j and $\hat{e}_Y(x_j)$ were divided by the number of the new and old form items before the computation; after summation, the value was multiplied by the number of the old form items to produce the final MAD statistic. The MAD values were compared to the "difference that matters (DTM)" criterion. The DTM value of .5 was used in this study, which is half of a score unit; a difference between equating results that is greater than the DTM is considered to be of practical significance (Dorans, Holland, Thayer, & Tateneni, 2003).

The test level evaluation was carried out for each of the FL and PHY tests by comparing equated raw scores with the identity equating results. Changes in profile shape were evaluated based on the relative rank order of subscores for each examinee. Specifically, in order to show whether an examinee's highest or lowest subscore changes before and after equating, the proportion of examinees in the new form group whose highest or lowest subscore changed was computed, respectively. The proportion of examinees whose overall subscore rank orders changed was also computed. These three proportion statistics (i.e., highest subscore change, lowest subscore change, and overall rank order change) show how equating affects examinees' subscore profiles. Note that because there are only two subtests for FL, the results for the three statistics should be the same. Profiles are also visually compared using graphs. The group mean profiles for the new form group are plotted before and after equating. In order to facilitate a comparison of raw subscores based on different numbers of items, they are transformed to proportion-correct scores before plotting.

3.2 Study 2: Simulated Data Analyses

Four factors were investigated in this study: test dimensionality, subtest length, form difference in difficulty, and examinee sample size. Test dimensionality was manipulated by adjusting correlation between ability parameters. The number of subtests is fixed at three, so each examinee has an ability vector, $\boldsymbol{\theta}$, whose length is three and each element of which corresponds to each subscore. Examinees' ability vectors were assumed to be multivariate normally distributed with a mean vector $\boldsymbol{\mu} = \mathbf{0}$ and a variance-covariance matrix

$$\begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

where ρ is correlation. In this study, three levels of correlation were employed: .7, .8, and .95. The number of items per subtest has four levels: 5, 10, 20, and 30. The number of subtests is fixed at three, so the total number of items is 15, 30, 60, and 90. The difference in difficulty between two forms was manipulated by changing the mean of the normal distribution from which MDIFF values were drawn. The mean difficulty was fixed at .00 for the old form, Form Y; for the new forms, Form X_1 and Form X_2, the mean difficulty values were fixed at .05, and .20. Thus, there are two levels in the form difference factor. Last, two different sample-size conditions were considered: 1,000 and 3,000. Thus, the total number of conditions is 48.

Item parameters for each form were generated based on approximate simple structure (APSS) using the multidimensional 3PL model (M3PL). Simulation conditions for item parameters were determined within realistic boundaries based on the M3PL item parameter estimates for the real data. For item j, $MDIFF_j$ was generated from a normal distribution with a variance of 1 and a mean set at one of the mean difficulty values specified above. $MDISC_j$ was generated using a beta distribution with $\alpha = 2$ and $\beta = 5$ and linearly transformed with slope=4.4 and intercept=.6 so that the mean and the standard deviation were set close to 1.9 and .7, respectively; these values were chosen based on the operational data used in Study 1. The correlation between $MDIFF$ and $MDISC$ was set to be between .15 and .25, which were also observed in the operational data analysis using M3PL. A NORTA (NORmal To Anything) transformation technique (Cario & Nelson, 1997; Yahav & Shmueli, 2012) was used to generate $MDIFF$ and $MDISC$ from different, yet somewhat correlated, distributions. The pseudo-guessing parameters c_j were generated from $beta(6, 16)$, and the intercept parameters d_j were computed using Equation 2:

$$MDIFF_j = \frac{-d_j}{\sqrt{\boldsymbol{a}_j \boldsymbol{a}_j'}} = \frac{-d_j}{MDISC_j}, \tag{2}$$

where $\boldsymbol{a}_j$ is a discrimination parameter vector. Generating the discrimination parameters a_{jk} based on APSS followed the procedures specified by Roussos, Stout, and Marden (1998).

After generating item parameters, ability vectors were generated from a multivariate normal distribution with the mean vector and variance-covariance matrix specified in the test dimensionality factor. In the third step, response patterns were simulated using the item and ability parameters. Then, 100 datasets were generated for each condition.

The equating methods employed in Study 1 were used in this study. The degree of presmoothing was determined based on the difference chi-square statistic. For the cubic-spline postsmoothing method, S=.01, .05, and .10 were considered which were most often selected in Study 1. From the results of Study 1, it was found that the three smoothing degrees produced similar results to each other with only minor differences. Thus, S=.05 was finally chosen in Study 2 to facilitate comparisons between equating methods.

To evaluate the performance of different equating methods, this study uses

the equipercentile equating results with the single group design using a large sample as the criterion equating relationships. Many studies have used population equipercentile equating to establish criterion equating relationships (Wang, Lee, Brennan, & Kolen, 2008). Ability vectors for 1,000,000 simulees were randomly drawn from a multivariate normal distribution with a mean vector and variance-covariance matrix specified in the test dimensionality factor. Number-correct subscores were simulated using the ability vectors and item parameters generated in the previous section. Equipercile equating was conducted and equating results were evaluated at each raw-score point using conditional bias (Bias), conditional standard error of equating (SEE), and conditional root mean square deviation (RMSD). These statistics were standardized by the standard deviation of the old form. The relative frequency distributions used to produce the criterion equating relationships were used to compute the old form standard deviation. Equations 3, 4, and 5 define the statistics:

$$Bias_j = \frac{\overline{\hat{e}_Y}(x_j) - e_Y(x_j)}{\sigma(Y)}, \tag{3}$$

$$SEE_j = \frac{\sqrt{\frac{1}{R}\Sigma_K\left[\hat{e}_Y(x_j) - \overline{\hat{e}_Y}(x_j)\right]}}{\sigma(Y)}, \tag{4}$$

and

$$RMSD_j = \sqrt{Bias_j^2 + SEE_j^2}, \tag{5}$$

where x_j is a raw-score point, $\hat{e}_Y(x_j)$ is an equated score at x_j, $\overline{\hat{e}_Y}(x_j)$ is the average of equated scores over replications, $e_Y(x_j)$ is the criterion equated score at x_j, $\sigma(Y)$ is the standard deviation of the old form, and R is the number of replications which was set to 100 in this study. Equating results were aggregated across all score points using weighted bias (WB), weighted standard error of equating (WSEE), and weighted root mean square deviation (WRMSD). The relative frequency of the new form from the criterion equating relationship was used to give weights to conditional results. For each subtest length condition, score points below the maximum for the sum of c-parameter estimates were not included in the computation. The overall statistics were defined as follows:

$$WB = \sqrt{\Sigma_j w_j Bias_j^2}, \tag{6}$$

$$WSEE = \sqrt{\Sigma_j w_j SEE_j^2}, \tag{7}$$

and

$$WRMSD = \sqrt{\Sigma_j w_j RMSD_j^2}, \tag{8}$$

where w_j is the relative frequency of x_j in the new form used to produce the criterion equating relationship.

4 Results

4.1 Operational Data Analyses

Tables 3 and 4 show that the overall differences between unequated and equated results represented by MAD statistics were different from subtest to subtest. It is evident that the overall differences from identity equating indicate form differences in difficulty that should be adjusted by conducting equating. For three subtests out of eight, MAD was greater than DTM no matter which equating method was employed, indicating that the average subscore changes after equating were practically significant.

Figures 1 and 2 contain difference plots for the FL and PHY subtests. The solid line in the middle represents identity equating, and the two dashed lines above and below the identity line represent DTM in the proportion-correct score unit. Differences between unequated and equated results conditional on new-form raw scores showed that subscore changes after equating might be greater than DTM in a specific score range even though the overall difference was smaller than DTM. There was only one subtest out of eight whose subscore changes after equating were smaller than DTM along the score scale. Although equating results for linear and IRT true score equating were somewhat different from other equating results, generally the ten equating methods were similar to each other in the difference plot, but not in identity equating. That is, the choice of whether or not to conduct equating at all appeared to be more fundamentally important than the choice of which equating method to use.

The subscore mean profiles for the new form group are presented in Figures 3 and 4. The equating methods yielded very similar mean equated scores although IRT true score equating results were slightly different from those of other methods. Thus, the shapes of profiles after equating were similar across the equating methods in the group mean profiles. For FL, which consists of two subtests, the shapes of group mean profiles did not change after equating. For PHY, which consists of six subtests, the shapes of group mean profiles changed after equating. No matter which equating method was applied, the rank orders of the subscore means were the same across the equating methods, which were different from the rank orders when equating was not conducted. The rank orders of the subscore means were 1) Subtest1 - Subtest3 - Subtest4 - Subtest5 - Subtest2 - Subtest6 before equating, and 2) Subtest1 - Subtest4 - Subtest3 - Subtest2 - Subtest6 - Subtest5 after equating. From these results, it is conceivable that there are many examinees whose individual-level score report for the relative performance across subtests changes depending on whether or not subscore equating is conducted.

Tables 5 and 6 provides the proportion of examinees whose subscore rank orders changed after equating. On average, more than 13% of the 3,000 examinees had different rank orders for the FL subtests after subtest form differences in difficulty were adjusted. For example, consider an examinee who scored 22 out of 30 (22/30=.7333) in the first subtest and 26 out of 35 (26/35=0.7429) in the second subtest. Because the proportion-correct score for the second sub-

test is higher, a score profile based on the unequated subscores indicates that this examinee performed better on listening (i.e., the second subtest). After equipercentile equating, however, the proportion-correct equated scores were .7833 for the first subtest and .7527 for the second subtest. That is, this examinee performed better on reading when subtest form differences in difficulty were adjusted through equating. The results for FL indicates that, even if the rank orders of group mean subscores remain the same before and after equating, individual level profiles can change. For PHY, equating changed the highest subscore for approximately 30% of examinees and the lowest subscore for approximately 25% of examinees. Moreover, about 92% of the examinees would receive a profile with different rank orders depending on whether equating was performed or not.

4.2 Simulated Data Analyses

Table 7 contains the overall summary statistics for the equating methods aggregated over all the conditions for Study 2. The WB values from Table 7 show that equipercentile with or without smoothing and 3PsepO produced smaller bias than did the other methods. Identity equating gave the largest WB among the methods. IRT true score equating methods tended to yield larger systematic error than the IRT observed score equating methods. With respect to WSEE, identity equating had no random error at all, and linear equating gave smaller random error than other methods. 3PsimO produced the third smallest WSEE, followed by BFO and 3PsepO. IRT observed score equating methods produced smaller WSEE than did equipercentile equating with or without equating. The average values of WSEE for the IRT true score equating methods were larger than the average values of WSEE for the other methods. The WRMSD values from Table 7 indicate that the best performing method is 3PsepO. Equipercentile with presmoothing showed a comparable performance to that of 3PsepO. The next best performing methods were 3PsimO and BFO. IRT true score equating and linear equating yielded the largest values of WRMSD among the methods. BFT resulted in the largest total error. Overall, IRT observed score equating and equipercentile equating with or without smoothing performed similarly in terms of total error. To be specific, equipercentile equating gave smaller bias and IRT observed score equating gave smaller random error compared to each other. Among the IRT observed score equating methods, 3PsepO produced the smallest bias and the largest random error. In contrast, 3PsimO produced the largest bias and the smallest random error. The performance of BFO lied in between 3PsepO and 3PsimO in both systematic and random errors. For IRT true score equating, the same pattern was observed as that of observed score equating: 3PsepT produced the smallest bias and the largest random error; 3PsimT gave the largest bias and the smallest random error; and BFT placed in between 3PsepT and 3PsimT in both bias and random error. Both presmoothing and postsmoothing reduced total error, with presmoothing performing better than postsmoothing.

Table 8 shows the overall summary statistics for the effects of test dimen-

sionality in Study 2. In general, the effect of test dimensionality on equating accuracy was not substantial and did not show a clear pattern, which was expected given that each subtest primarily measures only one dimension under the APSS assumption even if the total test has a multidimensional structure. The effect of test dimensionality on equating accuracy was not evident, partly because a narrow range of correlation coefficients was employed in the simulation—the correlation coefficients from .7 to .95 were selected to reflect a realistic range of multidimensionality. When IRT equating was conducted using 3PL with simultaneous calibration, systematic error tended to decrease as a test became more unidimensional. IRT true score equating using 3PL produced smaller random and total errors as a test became more unidimensional. Traditional equating and 3PsepO performed similarly across the test dimensionality conditions. With respect to test dimensionality, IRT equating using BF demonstrated a clear pattern. Bias, standard error and total error of equating after BFT or BFO were the smallest for the condition of ρ =.8, and larger for ρ =.7 and ρ =.95. Based on the mean and standard deviation of BF a_g and a_s estimates over replications, it was found that as a test became more unidimensional, a_g estimates were higher and more varied across replications, while a_s estimates were lower and less varied across replications. Because factor loading and variability for a_g and a_s change differently according to test dimensionality, equating error seems to reach its minimum or maximum at a modest level of test dimensionality (i.e., ρ =.8) rather than monotonically increasing or decreasing as the correlation coefficient changes.

With respect to subtest length, the overall summary statistics for Study 2 are presented in Table 9. In Study 2, the equating results for each subtest length condition were aggregated over only two alternate forms that differed in difficulty. As such, the effect of form characteristics confounded with the effect of subtest length could have a greater impact on the results, making it difficult to detect a clear pattern of equating error according to subtest length. Based on the results, it was found that unsmoothed equipercentile and smoothing methods tended to perform worse as subtest length increased in terms of random and total errors. On the other hand, IRT equating, except 3PsepO and BFO, performed better for a longer subtest. Presmoothing performed better in reducing random error as subtest length increased. Equipercentile equating with postsmoothing and IRT true score equating had a strong tendency to yield smaller systematic error as the subtest became longer.

Table 10 illustrates the overall summary statistics for Study 2 with respect to form difference in difficulty. When form difference in difficulty was larger, systematic and total errors of equating substantially increased, but random error of equating was not affected. Systematic and total error increased the most with linear and IRT true score equating. Under the small form difference condition, using BF yielded the smallest total error for both of IRT true and observed score equating. On the other hand, when form difference became larger, BFT and BFO gave larger total error than other IRT equating methods using 3PL.

Table 11 provides the overall summary statistics for the effects of sample size in Study 2. Not surprisingly, random error of equating decreased as sample size

increased, which led to a decrease in total error as well. With respect to random error of equating, equipercentile equating with or without smoothing was more affected by the change of sample size. It was also found that, compared to BF, IRT equating using 3PL benefited more from the increase of sample size. When sample size was 1,000, IRT equating using 3PL tended to produce larger random error than that for BFT or BFO; as sample size increased, random error for IRT equating using 3PL was smaller than that for BFT and BFO. These findings suggest that MIRT models including BF require larger sample size to reduce the standard error of equating to the same extent as for UIRT models. Similar patterns were observed in total error.

5 Discussion

This research provides some implications to be considered when subscores are reported as follows.

- Misleading information based on incorrect profiles
- Practical issues in form construction and equating
- The best performing method for subscore equating
- Borrowing information from other subtests
- MIRT equating for subscores
- Patterns of equating error when using BF
- Smoothing for subscore equating

First, reporting subscores without equating provides misleading information, not only due to the form difference of each subtest but also due to the relative form differences between multiple subtests. Subscores are differentiated from total scores in that the meaning of subscores is defined not only based on the corresponding subtest itself but also relative to the other subtests belonging to the same total test. Subscore equating is not supposed to preserve the profile patterns across forms, but the profile patterns can consequently be preserved when each subscore's meaning is sustained by conducting equating. Given that subscore reporting provides information on relative strengths and weaknesses in performance between subareas, incorrect profiles based on unequated subscores may lead to a huge waste of resources due to inappropriately directed investment. Second, reporting subscores involves consideration of a number of practical issues. For a very short subtest, constructing alternate forms with comparable difficulty is challenging, and equating results are not stable compared to those for longer subtests. When the number of items for a subtest differs across forms, equating subscores can be more complex. A detailed test specification is necessary for subtests so that each subtest has enough items to have meaning and to be equated adequately. Third, it appears that overall, 3PsepO

performs the best for subscore equating. Specifically, it performs the best when a test is almost unidimensional (ρ =.95), or subtest length is long (n =30). The second best method in terms of the overall performance is equipercentile with presmoothing. It performs the best when form difference is large, sample size is large (N =3,000), or subtest length is very short (n =5). Generally, IRT observed score equating performs the best followed by equipercentile equating with or without smoothing in terms of total error. IRT observed score equating produces larger bias but smaller random error compared to equipercentile equating with or without smoothing. The performance of linear equating and IRT true score equating is worse than the other methods. To be specific, BFT performs the worst among the equating methods. Fourth, for IRT true and observed score equating using 3PL, simultaneous calibration—which uses other subtest items in the calibration process—yields larger bias but smaller random error than does separate calibration, although they perform similarly with respect to total equating error. As shown in different contexts, such as subscore scoring and reporting, borrowing information from other subtests increases bias but decreases random error in subscore equating. Fifth, BFO tends to perform the best when a test is multidimensional (ρ =.7 and ρ =.8), form difference in difficulty is small, subtest length is short (n =10), or sample size is small (N =1,000). For both IRT true and observed score equating methods, using BF produces systematic error larger than that of 3PL with separate calibration but smaller than that of 3PL with simultaneous calibration; random error of BF equating is smaller than that of 3PL with separate calibration but larger than that of 3PL with simultaneous calibration. Sixth, IRT equating using BF yielded a different pattern of equating error than that of other equating methods, according to the test dimensionality conditions. Equating error was the smallest for the condition of ρ =.8, and larger for ρ =.7 and ρ =.95. This pattern may have occurred because the discrimination parameters of BF for the general and specific factors yielded opposite patterns in the mean and standard deviation over replications as a test became more unidimensional. Seventh, with respect to total equating error, smoothing improves equating results in general. However, postsmoothing introduced more bias than the amount of random error reduced when a subtest had only five items, the old and new forms differed substantially, and sample size was large.

There are several limitations that should be considered regarding the interpretation and generalization of the findings. First, the real data used in Study 1 were from tests that do not intend to report subscores. That is why the subtests examined in Study 1 have different numbers of items between subtests and even across forms. Moreover, it was assumed that subscores were reported based on only three or four items. There do exist operational testing programs that report subscores based on such a small number of items, but it would be a very extreme case for subtests belonging to the same total test to differ from 3 to 25 and for the number of subtest items to not be fixed across alternate forms. A study using real data from a test that reports subscores without equating should be conducted to generalize the findings in this research. Including a test that consists of distinct subareas such as reading, mathematics, and writing

would also be helpful to generalize the results. Second, the operational tests used in Study 1 are different with respect to how subscores are defined. The two subtests of FL are clearly distinguishable because each subtest consists of a contiguous set of items; in contrast, the six subtests of PHY consist of items that are interspersed throughout the test. In an operational test administration setting, PHY would involve much more significant context effects compared to FL. Such a difference was not reflected in the simulation studies in this paper. A simulation study could be designed in different ways by considering the context and position effects of subtest items. Third, a main focus of this research was on the performance of subscore equating methods that include other subtest items in the calibration process when estimating the item parameters for the target subtest. The four IRT equating methods with simultaneous calibration use items belonging to other subtests for calibration, but not for equating, suggesting that the results for the methods demonstrate bias and standard error of calibration rather than those of equating. Puhan and Liang (2011) suggested an equating method for the CINEG design that uses total scores as an anchor to produce more stable equating results. A comparison of this equating method against other methods under the CINEG design would provide a more direct presentation of bias and standard error of equating with respect to borrowing information from the total test. Fourth, this research employed M3PL to generate data, for which the number of dimensions was arbitrarily fixed at three. There is a possibility that the performances of IRT equating methods changes depending on the number of dimensions. Specifically, it is likely that BF estimating two orthogonal factors for a subtest might be more greatly affected by the number of dimensions. Peterson (2014) has already shown that BFO results in different equating relationships according to the number of dimensions specified. Including the number of dimensions as a study factor would provide in-depth information about how BFT and BFO perform in relation to test dimensionality.

Given that there exists such a strong demand for subscores, it would be hard for test developers to ignore it. It is hoped that the findings from this study provide test developers and users with some useful guidelines for selecting a method for subscore equating and urge them to consider various theoretical and practical issues with respect to subscore equating.

6 References

ACT. (2014). *Interpretive Guide for ACT Aspire Summative Reports* (Tech. Rep.). Retrieved from `http://actaspire.avocet.pearson.com/actaspire/Home#8439`

Brennan, R. L. (2011). *Utility indexes for decisions about subscores* (CASMA Research Report No. 33). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa.

Brennan, R. L., Toke, P., Kim, H. J., Lee, W., Kim, K. Y., & Lim, E. (2015). *Manual for ESUM-RG: Excel VBA Macros for Comparing Multiple Equat-*

ing Procedures under the Random Groups Design (CASMA Research Report No. 45). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa.

Brennan, R. L., Wang, T., Kim, S., & Seol, J. (2009). *Equating Recipes* (CASMA Monograph No. 1). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa.

Brossman, B. G., & Lee, W. (2013). Observed score and true score equating procedures for multidimensional item response theory. *Applied Psychological Measurement*, *37*(6), 460-481.

Cai, L. (2013). *flexMIRT Version 2: Flexible multilevel multidimensional item analysis and test scoring* [Computer Program]. Vector Psychometric Group.

Cario, M. C., & Nelson, B. L. (1997). *Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix* (Tech. Rep.). Evanston, IL: Department of Industrial Engineering and Management Sciences, Northwestern University.

College Board. (2014). *Test specifications for the redesigned SAT®*. Retrieved from `https://www.collegeboard.org/pdf/sat/delivering-opportunity/test_specifications_for_the_redesigned_sat_102414.pdf`

Dorans, N. J., Holland, P. W., Thayer, D. T., & Tateneni, K. (2003). Invariance of score linking across gender groups for three advanced placement program exams. In N. J. Dorans (Ed.), *Population invariance of score linking: Theory and applications to advanced placement program examinations* (p. 79-118). Princeton, NJ: Educational Testing Service.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*(2), 204-229.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1-28.

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

Lee, E. (2013). *Equating multidimensional tests under a random groups design: A comparison of various equating procedures* (Unpublished doctoral dissertation). The University of Iowa.

Lee, G., & Lee, W. (2016). Bi-factor MIRT observed-score equating for mixed-format tests. *Applied Measurement in Education*, *29*(3), 224-241.

Lee, G., Lee, W., Kolen, M. J., Park, I.-Y., Kim, D.-I., & Yang, J. S. (2015). Bi-factor mirt true-score equating for testlet-based tests. *Journal of Educational Evaluation*, *28*(2), 681-700.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement*, *8*(4), 453-461.

Peterson, J. L. (2014). *Multidimensional item response theory observed score equating methods for mixed-format tests* (Unpublished doctoral dissertation). The University of Iowa.

Puhan, G., & Liang, L. (2011). Equating subscores under the nonequivalent anchor test (neat) design. *Educational Measurement: Issues and Practice*, *30*(1), 23-35.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

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

Sinharay, S., & Haberman, S. J. (2011). Equating of augmented subscores. *Journal of Educational Measurement*, *48*(2), 122-145.

Stahl, J. A., & Masters, J. (2009). *Variable pass rates resulting from equating short tests.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Thorndike, R. M., & Thorndike-Christ, T. (2009). *Measurement and evaluation in psychology and education* (8th ed.). Upper Saddle River, NJ: Pearson.

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

Yahav, I., & Shmueli, G. (2012). On generating multivariate poisson data in management science applications. *Applied Stochastic Models in Business and Industry*, *28*(1), 91-102.

Table 1: Characteristics of FL and PHY Tests

	Test	Test Length New	Test Length Old	# of CIs	Sample Size (New, Old)
FL	Total	65	65	22	(17067, 14563)
	Reading	30	30	7	
	Listening	35	35	15	
PHY	Total	70	70	21	(17625, 60393)
	Classical Mechanics	22	25	6	
	Electricity & Magnetism	20	16	6	
	Waves & Optics	14	11	4	
	Thermal Physics	5	8	2	
	Fluid Mechanics	4	3	0	
	Atomic & Nuclear Physics	5	7	3	

Table 2: Degree of Smoothing for Subtests

Test	Subtest	Log-Linear Presmoothing		Cubic-Spline
		New Form	Old Form	Postsmoothing
FL	Reading	CX = 4	CY = 5	S = .05
	Listening	CX = 6	CY = 8	S = .05
PHY	CM	CX = 4	CY = 6	S = .10
	EM	CX = 4	CY = 4	S = .05
	WO	CX = 4	CY = 6	S = .05
	TP	CX = 4	CY = 4	S = .01
	FM	CX = 4	CY = 3	S = .01
	AN	CX = 4	CY = 5	S = .01

Note: CM=Classical Mechanics; EM=Electricity & Magnetism; WO=Waves & Optics; TP=Thermal Physics; FM=Fluid Mechanics; AN=Atomic & Nuclear Physics.

Table 3: MAD for FL Subtests

Method	Subtest	
	Reading	Listening
Linear	1.2253	0.3552
Equipercentile	1.2323	0.4834
Presmoothing	1.2291	0.4815
Postsmoothing	1.2311	0.4862
3PsepT	1.1628	0.4748
3PsepO	1.2234	0.4460
3PsimT	1.1656	0.4351
3PsimO	1.2190	0.4100
BFT	1.1523	0.3955
BFO	1.2302	0.4459
Average	1.2071	0.4414

Table 4: MAD for PHY Subtests

Method	Subtest					
	CM	EM	WO	TP	FM	AN
Linear	0.4395	0.2741	0.8302	0.8667	0.0941	0.4923
Equipercentile	0.4418	0.3015	0.8325	0.8787	0.1068	0.4978
Presmoothing	0.4408	0.2915	0.8336	0.8826	0.1068	0.4969
Postsmoothing	0.4335	0.3019	0.8327	0.8788	0.1021	0.5038
3PsepT	0.4267	0.2644	0.6943	0.6537	0.0746	0.3630
3PsepO	0.4351	0.2789	0.7710	0.8784	0.1045	0.4965
3PsimT	0.4187	0.2079	0.7436	0.6291	0.0837	0.3495
3PsimO	0.4374	0.2395	0.8289	0.8866	0.0936	0.4893
BFT	0.4141	0.2243	0.7430	0.6256	0.0829	0.3355
BFO	0.4434	0.2556	0.8386	0.8894	0.0995	0.4919
Average	0.4331	0.2640	0.7948	0.8070	0.0949	0.4517

Table 5: Proportion of Examinees Whose Rank Orders Change for FL

Method	Rank Order Change
Linear	0.1467
Equipercentile	0.1303
Presmoothing	0.1177
Postsmoothing	0.1177
3PsepT	0.1263
3PsepO	0.1337
3PsimT	0.1267
3PsimO	0.1600
BFT	0.1473
BFO	0.1417
Average	0.1348

Table 6: Proportion of Examinees Whose Rank Orders Change for PHY

Method	Highest Subscore Change	Lowest Subscore Change	Overall Rank Order Change
Linear	0.3183	0.2337	0.9350
Equipercentile	0.3427	0.2513	0.9437
Presmoothing	0.3430	0.2510	0.9450
Postsmoothing	0.3420	0.2540	0.9447
3PsepT	0.2660	0.2443	0.8903
3PsepO	0.3383	0.2297	0.9360
3PsimT	0.1803	0.2533	0.8780
3PsimO	0.3420	0.2490	0.9410
BFT	0.1810	0.2510	0.8767
BFO	0.3417	0.2443	0.9433
Average	0.2995	0.2462	0.9234

Table 7: Summary Statistics for Equating Methods in Study 2

Method	WB	WSEE	WRMSD
Identity	0.2047	0.0000	0.2047
Linear	0.0500	0.0391	0.0652
Equipercentile	0.0046	0.0465	0.0467
Presmoothing	0.0058	0.0452	0.0456
Postsmoothing	0.0112	0.0443	0.0465
3PsepT	0.0400	0.0481	0.0650
3PsepO	0.0113	0.0430	0.0450
3PsimT	0.0474	0.0452	0.0670
3PsimO	0.0209	0.0399	0.0460
BFT	0.0474	0.0464	0.0676
BFO	0.0176	0.0413	0.0460

Table 8: Summary Statistics under the Condition of Test Dimensionality in Study 2

	WB			WSEE			WRMSD		
Method	ρ=0.7	ρ=0.8	ρ=0.95	ρ=0.7	ρ=0.8	ρ=0.95	ρ=0.7	ρ=0.8	ρ=0.95
Identity	0.2057	0.2046	0.2038	0.0000	0.0000	0.0000	0.2057	0.2046	0.2038
Linear	0.0500	0.0495	0.0504	0.0394	0.0390	0.0390	0.0655	0.0647	0.0653
Equipercentile	0.0051	0.0044	0.0044	0.0468	0.0461	0.0464	0.0471	0.0464	0.0467
Presmoothing	0.0063	0.0050	0.0059	0.0457	0.0447	0.0452	0.0461	0.0451	0.0457
Postsmoothing	0.0115	0.0110	0.0112	0.0447	0.0439	0.0443	0.0470	0.0459	0.0465
3PsepT	0.0411	0.0391	0.0398	0.0490	0.0482	0.0469	0.0666	0.0650	0.0634
3PsepO	0.0119	0.0106	0.0113	0.0436	0.0434	0.0420	0.0456	0.0452	0.0442
3PsimT	0.0515	0.0457	0.0450	0.0468	0.0459	0.0429	0.0711	0.0661	0.0637
3PsimO	0.0238	0.0193	0.0195	0.0403	0.0409	0.0384	0.0474	0.0462	0.0442
BFT	0.0500	0.0458	0.0463	0.0472	0.0453	0.0467	0.0700	0.0658	0.0671
BFO	0.0170	0.0158	0.0201	0.0402	0.0403	0.0435	0.0443	0.0444	0.0493
Average	0.0325	0.0294	0.0303	0.0445	0.0440	0.0434	0.0575	0.0554	0.0553

Table 9: Summary Statistics under the Condition of Subtest Length in Study 2

Method	WB				WSEE				WRMSD			
	n=5	n=10	n=20	n=30	n=5	n=10	n=20	n=30	n=5	n=10	n=20	n=30
Identity	0.2859	0.1617	0.2279	0.1433	0.0000	0.0000	0.0000	0.0000	0.2859	0.1617	0.2279	0.1433
Linear	0.0516	0.0465	0.0606	0.0412	0.0389	0.0388	0.0400	0.0388	0.0660	0.0628	0.0747	0.0572
Equipercentile	0.0035	0.0045	0.006	0.0045	0.0428	0.0459	0.0487	0.0485	0.0429	0.0461	0.0491	0.0488
Presmoothing	0.0044	0.0058	0.0071	0.0058	0.0423	0.0449	0.0471	0.0465	0.0426	0.0453	0.0477	0.0470
Postsmoothing	0.0238	0.0084	0.0077	0.0051	0.0418	0.0431	0.0463	0.0460	0.0485	0.0439	0.0470	0.0464
3PsepT	0.0755	0.0370	0.0274	0.0200	0.0521	0.0483	0.0486	0.0432	0.0925	0.0624	0.0573	0.0477
3PsepO	0.0062	0.0066	0.0171	0.0152	0.0428	0.0419	0.0457	0.0417	0.0433	0.0427	0.0494	0.0446
3PsimT	0.0750	0.0471	0.0376	0.0299	0.0500	0.0476	0.0431	0.0401	0.0908	0.0677	0.0591	0.0503
3PsimO	0.0116	0.0168	0.0293	0.0258	0.0416	0.0402	0.0394	0.0382	0.0433	0.0439	0.0502	0.0465
BFT	0.0732	0.0387	0.0428	0.0348	0.0531	0.0484	0.0446	0.0395	0.0912	0.0629	0.0634	0.0529
BFO	0.0103	0.0105	0.0254	0.0243	0.0451	0.0412	0.0414	0.0377	0.0464	0.0426	0.0497	0.0453
Average	0.0335	0.0222	0.0261	0.0207	0.0451	0.0440	0.0445	0.0420	0.0608	0.0520	0.0548	0.0487

Table 10: Summary Statistics under the Condition of Form Difference in Difficulty in Study 2

Method	WB		WSEE		WRMSD	
	Diff.=.05	Diff.=.20	Diff.=.05	Diff.=.20	Diff.=.05	Diff.=.20
Identity	0.1185	0.2909	0.0000	0.0000	0.1185	0.2909
Linear	0.0327	0.0673	0.0390	0.0393	0.0516	0.0788
Equip.	0.0044	0.0048	0.0474	0.0455	0.0477	0.0458
Pre.	0.0051	0.0064	0.0461	0.0443	0.0464	0.0448
Post.	0.0090	0.0134	0.0443	0.0443	0.0456	0.0474
3PsepT	0.0296	0.0503	0.0478	0.0483	0.0583	0.0717
3PsepO	0.0079	0.0146	0.0438	0.0423	0.0447	0.0453
3PsimT	0.0366	0.0582	0.0449	0.0455	0.0591	0.0748
3PsimO	0.0168	0.0250	0.0407	0.0391	0.0443	0.0476
BFT	0.0336	0.0612	0.0455	0.0473	0.0573	0.0779
BFO	0.0131	0.0221	0.0414	0.0413	0.0437	0.0483
Average	0.0189	0.0323	0.0441	0.0437	0.0499	0.0582

Table 11: Summary Statistics under the Condition of Sample Size in Study 2

Method	WB		WSEE		WRMSD	
	N=1,000	N=3,000	N=1,000	N=3,000	N=1,000	N=3,000
Identity	0.2047	0.2047	0.0000	0.0000	0.2047	0.2047
Linear	0.0504	0.0496	0.0501	0.0282	0.0725	0.0579
Equip.	0.006	0.0032	0.0593	0.0336	0.0597	0.0338
Pre.	0.0074	0.0042	0.0576	0.0328	0.0581	0.0332
Post.	0.0125	0.0100	0.0564	0.0322	0.0584	0.0345
3PsepT	0.0400	0.0400	0.0576	0.0385	0.0725	0.0575
3PsepO	0.0150	0.0075	0.0524	0.0337	0.0552	0.0348
3PsimT	0.0500	0.0448	0.0549	0.0355	0.0756	0.0583
3PsimO	0.0259	0.0158	0.0492	0.0306	0.0567	0.0353
BFT	0.0507	0.0441	0.0550	0.0379	0.0762	0.0591
BFO	0.0229	0.0123	0.0487	0.0340	0.0551	0.0369
Average	0.0281	0.0231	0.0541	0.0337	0.0640	0.0441

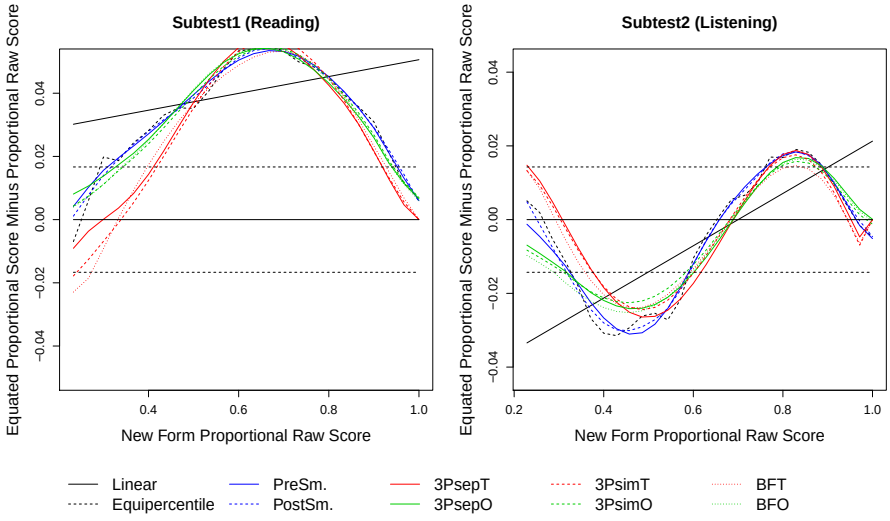

Figure 1: Difference Plot for FL Subtests

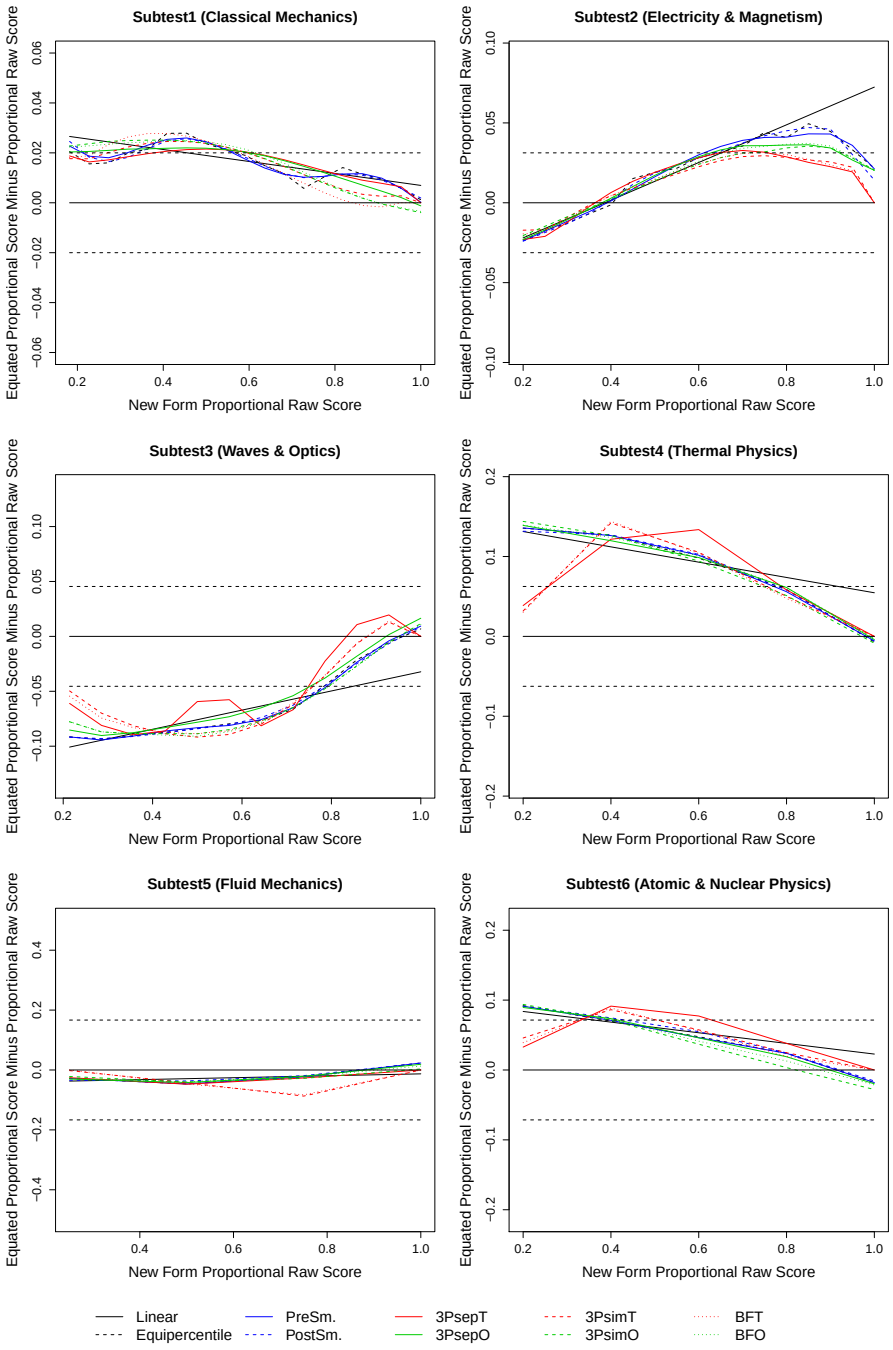

Figure 2: Difference Plot for PHY Subtests

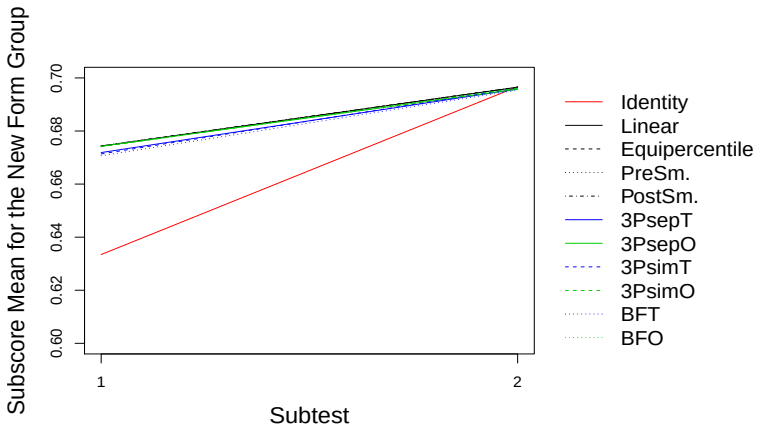

Figure 3: FL Group Mean Profile for the New Form Group

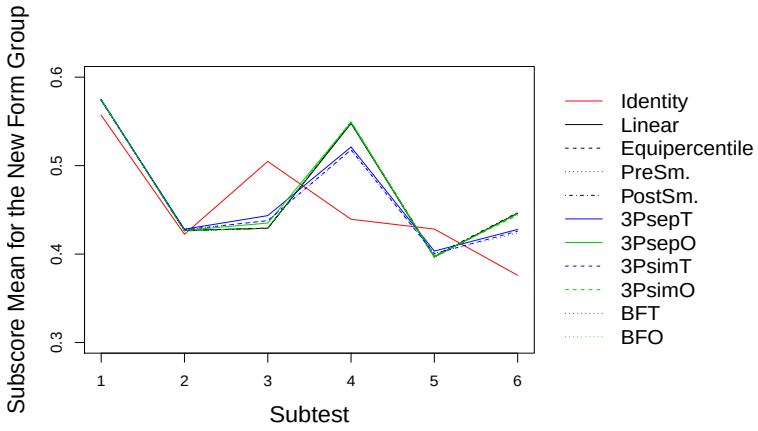

Figure 4: PHY Group Mean Profile for the New Form Group