



Listening. Learning. Leading.

Repeater Effects on Score Equating for a Graduate Admissions Exam

Wen-Ling Yang

Andrea M. Bontya

Tim P. Moses

ETS, Princeton, NJ

Paper presented at the annual meeting of the
American Educational Research Association (AERA)

April 13-17, 2009, San Diego, CA.

Unpublished Work Copyright © 2009 by Educational Testing Service. All Rights Reserved. These materials are an unpublished, proprietary work of ETS. Any limited distribution shall not constitute publication. This work may not be reproduced or distributed to third parties without ETS's prior written consent. Submit all requests through www.ets.org/legal/index.html.

Educational Testing Service, ETS, the ETS logo, and *Listening. Learning. Leading.* are registered trademarks of Educational Testing Service (ETS).



Abstract

With self-reported but empirically verified repeater groups, we analyzed vast amount of real test data across a wide range of administrations from a graduate admissions examination that was administered in a non-English language to investigate repeater effects on score equating using the non-equivalent groups anchor test (NEAT) design. Both linear and non-linear equating models were considered in deriving the equating functions for various study groups. We evaluated scaled score differences between equatings in the total group, the repeater group and the first-timer group using statistics of simple differences and subpopulation invariance measures developed and used widely in the last 10 years. Standard errors of statistics summarizing scaled score differences were estimated using a simulation approach to provide statistical criteria for evaluating the significance of equating differences. In addition, we used scaled score differences that were critical to admissions screening as criteria for evaluating practical significance of equating differences. To put the investigation of repeater effects in proper perspective, we analyzed the repeater data for an in-depth understanding of repeater performance trends. Overall, we found no significant effects of repeater performance on score equating for the study exam. Although many of the equating differences were practically significant, most of the practically significant differences were not statistically significant. However, further research with larger repeater samples was recommended to help explain the practical significance of equating differences consistently observed in this study for the repeater group. Potential problems associated with small repeater study sample sizes, issues with the practical criterion for evaluating the significance of equating differences and study limitations were also discussed.



Introduction

Score equating is commonly used for ensuring comparable scores across different test administrations and/or forms. A variety of equating methods have been developed and used in practice and these methods have been well researched under a broad range of conditions, such as characteristics of test/anchor/sample and mix of content or item format. However, little attention has been given to potential repeater effects on score equating, which is especially important for testing programs with a high percentage of repeating examinees. The effect of repeaters and equating could influence each other in a reciprocal way. As choices of equating design and sample treatment should take into account repeater effects, evaluation of the repeater effects based on equated scores, such as repeater gain or loss, will depend on equating outcomes.

Although some testing programs from time to time review repeater rates and patterns, repeater effects are usually evaluated in the context of scale score gain/loss across test administrations and interventions are seldom in place to directly address potential effects of repeater performance on equating outcomes, which could introduce bias in the estimation of ability distributions for equating. Even for programs that routinely exclude repeaters from the equating process to control the potential systematic bias due to the repeater performance, there is often a lack of evaluation of repeater effects on equating such that it is not certain whether this practice of excluding repeaters is appropriate in terms of fairness or for ensuring equating quality. As equating is generally more adequate when the examinees included in the equating samples are as similar as possible to the entire group tested (Harris, 1993), by excluding repeaters (especially for a large repeater group) an equating sample may become smaller in size and/or less representative of the total examinee group, which may have a negative impact on equating precision (Kolen & Brennan, 1995). Thus, a concern naturally rises over the practice of excluding repeaters from equating process, especially when the direction and magnitude of the repeater effects are not clear.

Previous research about repeater effects generally focused on studying score stability over testing occasions, forms, formats, and/or modalities (Zhang, 2008; Gorham & Bontempo, 1996; Kingston & Turner, 1984). And, changes in scale scores, ability



estimates, and/or passing rates were often the unit of analysis, despite the fact that equating was critical in deriving the scale scores and ability estimates and in determining the passing rates. Only a few research studies directly investigated the effect of repeaters in the context of score equating. The case study of Andrulis, Starr, and Furst (1978) published more than 30 years ago was a pioneer in this area, which examined the impact of repeater performance on a linear equating model based on the random equivalent groups design with anchor items and evaluated the repeater effects in terms of differences in the resulting equating parameters, cut-off points, and passing rates. The authors found the self-selected repeaters in this case study to be less able than the first-time examinees and the performance of the less able repeaters contributed to a lowered passing score. As a practical solution to meet the equating assumption (i.e., random groups) and to mitigate the repeater effects, the authors suggested the removal of repeaters from the process of deriving an equating conversion. Another example of the research in this area is the equating study by Cope (1986) based on the non-equivalent groups anchor item (NEAT) design. In his study, Cope compared the results of linear equating models using examinee data with and without repeaters and used the equating chains to evaluate the relative accuracy of various equatings. Because the equatings based on the first-time examinees were not necessarily/substantially more accurate than the equatings based on the total examinee group, and the relative accuracy of equating seemed to depend partly on the specific linear equating method used, the author had reservations about the practice of routinely excluding repeaters from equating for the test being studied and called for more research to investigate whether the equating differences would become larger when there was a larger repeater group.

In summary, the effects of repeaters on equating seem to be dependent of the size and ability of the repeater group, which are under the influence of the other characteristics of the repeater group (e.g., motivation and preparation levels), the purpose and use of the test (e.g., low vs. high risk, with vs. without a threshold), as well as the test characteristics (e.g., content is subject to practice effect or not). And, the repeater effects on equating may also depend on the equating design and method used. As a result, the repeater effects are likely to be test specific and can vary widely across testing programs. So far, the limited number of equating research studies that focused on the repeater



effects had used data from different testing programs and involved different equating designs and methods, and the results looked mixed and may not be generalized to the equatings for other testing conditions. There is clearly a need for more research to expand our knowledge about repeater effects on equating to ensure equating accuracy and test fairness, even if the study has to be conducted on a case by case basis. Hopefully, by accumulating a wealth of systematic empirical research results we will be able to better delineate the effects of repeater performance on score equating and to prescribe adequate strategies for handling the equatings in various testing conditions.

Therefore, using real test data from multiple test administrations of an operational examination that was administered in a non-English language we investigated the repeater effect on score equating under the following conditions:

- Test use/purpose—Graduate admissions of medium to high stake.
- Ability measures—General skills required by graduate studies.
- Primary definition of repeaters—Examinees that repeated the exam at least once, regardless of the time interval between testing occasions and/or the number of retakes. In other words, the repeater group analyzed in this study was primarily a sample of the overall, non-specific repeater population, unless otherwise specified.
- Repeater identification—Self-reported but empirically verified.
- Repeater group—Fairly large in size and more able than the first-time examinee group on average.
- Study data—Real test data from multiple exam administrations for studying the repeater patterns and effects, and simulated data based on the real target equating samples for estimating the standard errors of equating (SEEs) and the standard errors of equating differences (SEEDs).
- Equating design—Non-equivalent groups anchor test (NEAT) design.
- Equating methods—Both linear and non-linear models.
- Unit of analysis—Equating outcomes expressed on the reporting scale (i.e., scale scores), instead of the raw-score scale.



- Tools for summarizing the repeater effects—Multiple summary statistics for describing the equating differences between the subgroups and between the total group and individual subgroups.
- Criteria for evaluating the significance of the repeater effects—Both statistical and practical evaluation criteria.

Objectives

Primary objectives of this study are as follows:

- (1) To assess repeater effects on score equating and evaluate their statistical and practical significance.
- (2) To discuss the implications of repeater effects on scoring fairness and make recommendations about the treatment/use of repeater data for equating, especially for testing programs that deal with a significant number of repeaters.

This study also has the following secondary objectives that are specific to the testing program being studied:

- (3) To delineate the general patterns of the repeater rate/trend for the study exam and to evaluate the soundness of the program's operational practice of excluding repeaters from the equating samples, which should provide useful information for the testing program in improving its equating procedure and ensuring score fairness, and help the test users to interpret multiple pieces of testing outcomes resulting from different testing occasions for the same examinee.
- (4) To verify the self-reported repeater data using the empirical test-taking information across exam administrations and to evaluate the validity of the survey question used by the study exam that asked the examinees to identify their repeater status on a voluntary basis.



Data

To ensure representative and stable analysis outcomes, in this study we used real multiple-administrations test data from an examination that was administered in a non-English language and primarily used for making decisions about graduate admissions and granting scholarships. Consisting of all multiple-choice items in the paper-and-pencil test format, the study exam measures four general skills required for graduate studies and the test consequences are of medium to high stakes. The testing program currently permits examinees to take the exam multiple times without any limit on the number of retakes or the time interval between test and retest. Despite the program policy that holds the reported scale scores valid for five years, examinees (even those who scored high previously) have an incentive to retake the exam sooner to achieve a higher score to increase their chance of being admitted to an institution with higher admission standards. Across the exam administrations, the self-reported repeater rate ranged from about 20% to 40%. The program routinely excludes self-reported repeaters from equating based on the assumption that the repeaters would have an advantage over the first-time examinees due to the practice effect, while scores on different test forms are equated operationally using the non-equivalent groups anchor test (NEAT) design.

For the sake of practical and feasible research scope, we conducted an in-depth investigation focusing on the two core skills measured by the exam (specifically, the verbal and quantitative measures). Close to seven years worth of recent operational data from 2000 to 2007 were analyzed for various study purposes, which included data on the targeted new and reference forms from multiple exam administrations, as well as aggregated data over five consecutive years prior to each of the targeted new form administrations in order to retrieve sufficient examinee records for verifying the self-reported repeater status and analyzing the repeater rates/patterns. The decision to backtrack five years worth of score data was based on the assumption that an examinee who took the exam more than five years ago was less likely to repeat the study exam and if the examinee did repeat the exam he/she might not benefit significantly from the prior test-taking experience.



In addition to the real test data described above, we also used simulated data based on the real equating samples to estimate the standard errors of equatings (SEEs) and the standard errors of equating differences (SEEDs). While we will describe the simulation approach in detail later in the method section, we summarize the characteristics of the real test data used for the study analyses below.

We analyzed data from three test administrations for the Quantitative measure and data from two administrations were analyzed for the Verbal measure¹. In general, there were 65 items in a Quantitative test form and 90 items in a Verbal form. However, the actual test length varied due to the removal of items with poor performance before equating/scoring. For each of the forms analyzed in this study as “new” forms, Table 1 shows the possible score ranges for the total test and the anchor test, respectively, as well as the sample sizes for the first-time examinee group, the repeater group, and the total group. For each of the new-form examinee groups, Table 1 also presents the means and standard deviations of scores on the total test and the anchor test, and the coefficient of correlation between the total and anchor test scores. Table 2 presents similar information for study forms analyzed as reference forms. The “-V” and “-Q” in the form names indicate whether a test form is for the Verbal or Quantitative measure.

¹ Ideally, we would like to include data from three test administrations for the Verbal measure, like what we had done for the Quantitative measure. However, we only used data from two administrations due to data availability.



Table 1. Summary Statistics for New Forms by Examinee Group

New Form	Possible Score Range		Examinee Group	N	Test Score		Anchor Score		Anchor-Total Correlation
	Total Test	Anchor Test			Mean as % of possible max.	SD as % of possible max.	Mean as % of possible max.	SD as % of possible max.	
A-V	0-86	0-25	1st-timer	1,419 (73%)	53.9%	12.5%	53.4%	16.2%	0.87
			Repeater	537 (27%)	57.5%	11.5%	57.7%	15.0%	0.86
			Total	1,956	54.9%	12.4%	54.6%	16.0%	0.87
A-Q	0-64	0-28	1st-timer	1,419 (73%)	45.4%	15.9%	44.6%	19.3%	0.93
			Repeater	537 (27%)	47.9%	14.8%	48.6%	18.3%	0.93
			Total	1,956	46.1%	15.7%	45.7%	19.1%	0.93
B-V	0-86	0-24	1st-timer	989 (79%)	60.6%	14.2%	54.1%	17.6%	0.85
			Repeater	261 (21%)	63.8%	12.8%	58.5%	16.0%	0.83
			Total	1,250	61.2%	14.0%	55.0%	17.4%	0.85
B-Q	0-62	0-20	1st-timer	989 (79%)	37.0%	11.7%	35.8%	13.6%	0.79
			Repeater	261 (21%)	38.4%	11.0%	37.4%	13.1%	0.77
			Total	1,250	37.3%	11.6%	36.1%	13.5%	0.79
C-Q	0-65	0-18	1st-timer	1,234 (72%)	46.1%	16.6%	57.8%	20.3%	0.87
			Repeater	474 (28%)	46.6%	15.2%	58.9%	19.3%	0.85
			Total	1,708	46.2%	16.2%	58.2%	20.0%	0.86

Table 2. Summary Statistics for Reference Forms by Examinee Group

Reference Form	Possible Score Range		Examinee Group	N	Test Score		Anchor Score		Anchor-Total Correlation
	Total Test	Anchor Test			Mean as % of possible max.	SD as % of possible max.	Mean as % of possible max.	SD as % of possible max.	
RA-V	0-88	0-25	1st-timer	1,239 (65%)	55.9%	12.5%	55.0%	16.8%	0.89
			Repeater	653 (35%)	58.7%	11.7%	58.2%	15.3%	0.86
			Total	1,892	56.9%	12.3%	56.1%	16.4%	0.88
RA-Q	0-65	0-28	1st-timer	1,178 (72%)	47.4%	17.3%	45.7%	20.6%	0.93
			Repeater	453 (28%)	48.2%	15.3%	47.7%	18.6%	0.92
			Total	1,631	47.6%	16.8%	46.2%	20.1%	0.93
RB-V	0-84	0-24	1st-timer	1,081 (78%)	53.9%	13.9%	52.0%	16.7%	0.84
			Repeater	312 (22%)	56.6%	13.5%	55.0%	15.9%	0.83
			Total	1,393	54.5%	13.9%	52.7%	16.5%	0.84
RB-Q	0-64	0-20	1st-timer	1,081 (78%)	38.3%	11.3%	37.0%	14.4%	0.81
			Repeater	312 (22%)	38.0%	10.2%	36.1%	13.1%	0.73
			Total	1,393	38.2%	11.0%	36.8%	14.2%	0.80
RC-Q	0-65	0-18	1st-timer	1,753 (76%)	59.3%	17.7%	60.5%	20.6%	0.90
			Repeater	554 (24%)	58.7%	16.8%	60.4%	19.7%	0.88
			Total	2,307	59.2%	17.5%	60.5%	20.4%	0.90



Tables 1 and 2 show that the percentage of repeaters across the ten study forms ranged from 21% to 35%, which was fairly consistent with the percentage range based on all of the available study data across a larger number of forms/administrations. Overall, the tables show that the mean scores (on both the total test and the anchor test) of the repeater group were consistently higher than those for the first-time examinee group across various forms, except for two Quantitative reference forms (namely, RB-Q and RC-Q), and in general the repeater group was less variable than the first-time examinee group. We will present in detail the group comparison outcomes in the Results section, following a description of the methods used for the group comparisons in the Method section.

Also shown in Tables 1 and 2 was that the anchor-total correlation coefficients ranged widely from 0.73 to 0.93 across test forms and examinee groups. While the variation in the correlation coefficient looked quite large across study forms, the variation across examinee groups was actually quite smaller except for Form RB-Q (for which the anchor-total correlation for the repeater group was much lower than those for the total group and the first-time examinee group). Such variation in the anchor-total correlation might lead to different levels of equating efficacy across various forms (and, across the examinee groups for Form RB-Q only).

A close inspection on the raw mean scores (as the percentage of possible maximum score points) presented in Tables 1 and 2 also indicated that means across Verbal forms were less variable than the Quantitative means for both the total and anchor tests. Although differences in test difficulty across forms could not be determined until scores on different test forms were equated, it was possible that the Verbal forms constructed were more comparable to each other than the Quantitative forms. In addition, since group differences in ability could also contribute to the variation across administrations in raw mean scores, differences between the Verbal and Quantitative score data might imply that the examinee groups across administrations overall possessed similar levels of knowledge/skills on Verbal but not on Quantitative. This implication sounds reasonable because Verbal and Quantitative forms measured very different constructs; and, as a result, examinee groups across administrations that performed similarly on one measure might not perform as similarly on the other measure.



Method

In this section we first define the repeaters for this study and summarize the approaches used for identifying and verifying the repeater information. Then, we will describe the methods used for analyzing general repeater trends, followed by the methods for investigating the effects of repeater performance on equating and for evaluating the statistical and practical significance of the repeater effects. By analyzing the general trends of the repeater group and their performance, we can put the investigation of the repeater effects on equating in proper perspective.

Definition of Repeaters

The repeaters in this study are defined as examinees that repeated the exam at least once, regardless of the time interval between testing occasions and/or the number of retakes, unless otherwise specified. In other words, the repeater group for the equating study was a sample of the overall non-specific repeater population. Our study samples were not very large to begin with, so the further breakdowns of the study samples by specific repeater characteristics such as the number of retakes (e.g., the 1st-time repeaters, the 2nd-time repeaters, the 3rd-time repeaters, and so on) would not support meaningful statistical outcomes. For example, the average size of the repeater groups that repeated the exam two or more times could be as small as 100, not larger than 200, for the target study forms. As indicated by Gorham and Bontempo (1996), inferences based on the repeater subpopulations characterized by the number of retakes were likely to be unstable because the amount of data dwindled quickly across retests. The amount of data could also decrease dramatically when the repeater group was broken down by the other characteristics, such as the time interval between test and retest (e.g., within six months, one year, two years, etc.) and the ability levels of repeaters. Therefore, in this study we focused on examining the effects of the overall, non-specific repeater group, instead of the effects of any specific repeater subgroups.

Identification and Verification of Repeater Status

The repeaters in this study were identified based on examinees' voluntary responses to a survey question that asked them whether they were retaking the exam. Because the



examinees were more likely to disguise their repeater status than to identify themselves as repeaters when they were actually not (there was an incentive to alienate themselves from the previous record of poor performance), the self-reported repeater status may not accurately reflect examinees' true repeater status. As a measure to verify the accuracy of the self-reported repeater status, we compared the self-reported repeater data to the empirically identified repeater data, which was derived by matching examinee records in the study database across multiple exam administrations using available identifying information such as the social security numbers, names, addresses, birth dates, etc.

General Trend of Repeater Performance

For a general understanding of the overall trend of repeater performance on the exam, we computed the test-retest correlation coefficient using scale score data in the repeater group, investigated examinees' scale score gain/loss after retaking the exam, and compared the repeater group's performance to the performance of first-time examinee group on the raw total test scale.

Test-Retest Correlation

Prior to the consideration of repeater effects on score equating, we calculated the test-retest correlation coefficient to show the relationship between the test and the retest scores of the same examinee group obtained from two different but consecutive testing occasions in the context of equated scores. Specifically, we correlated the scale scores from the two most recent testing occasions for the overall repeater group. Since some of the examinees had repeated the exam more than once, we focused on the two most recent testing occasions for each of the repeaters in this analysis to standardize the selection of test scores and to take into account data recency. Because examinees in the overall repeater group came from a wide range of administrations, the most recent test (or the second most recent test) taken by different repeaters might not be the same.

To study whether the test-retest relationship depended on the distance in time between the two testing occasions, we also computed the test-retest correlation coefficients for repeater subgroups that differed in test-retest interval time.



Scale Score Gain/Loss

We examined the general scale score gain/loss for the overall, non-specific repeater group. To acquire an in-depth understanding of the repeater performance trend, we further compared the patterns of scale score gain/loss across various repeater subgroups that differed in their performance on the previous testing occasion. Specifically, conditional distribution data for the repeater gain/loss (i.e., the percentages of repeaters with scale score gain/loss on the most recent testing occasion conditioned on their scale scores on the previous testing occasion) was used for this analysis. The conditional distribution results should provide more detailed information regarding repeater gain/loss, while taking into account the repeaters' prior performance.

In addition, to study the patterns of repeater score gain/loss in a more restrictive context (i.e., restricting the repeater group within a more specific timeframe) we analyzed the much smaller sample of data for the repeater group taking the most recent study test. By focusing on the performance of a more narrowly defined repeater group, we could mitigate the potential regressive effects caused by aggregating repeater groups across administrations.

Comparing repeater performance to first-timer performance

To study repeater performance that was not influenced by the equating practice and its effects on scoring consequences, for each of the target equating forms we also compared the overall performance of the repeater group to the performance of the first-time examinee group using their raw total scores on the same test form². Specifically, we plotted the observed relative frequency distributions between the repeater and the first-time examinee groups on the raw total score scale to show how the two groups differed as a whole. We also inspected the mean score differences between the two groups and evaluated the statistical significance of the group differences by using the two-samples Z test as follows:

² We could have also compared the repeater group's performance to the first-time examinee group's performance by using their scores on the same anchor test. However, we decided to focus the comparison on the (raw) total test scores because they were more reliable and representative of the examinees' performance than the anchor test scores.



$$Z = \frac{(\bar{X}_r - \bar{X}_{nr}) - (\mu_r - \mu_{nr})}{\sqrt{\sigma_{\bar{X}_r}^2 + \sigma_{\bar{X}_{nr}}^2}},$$

where $\sigma_{\bar{X}_r}^2 = \frac{\sigma_r^2}{n_r}$ and $\sigma_{\bar{X}_{nr}}^2 = \frac{\sigma_{nr}^2}{n_{nr}}$.

Repeater Effects on Score Equating

To examine the repeater effects on equating, we compared the equating function derived using the first-time examinee group data to the function based on the total group data and evaluated the significance of equating differences using both the statistical and practical criteria. We also compared the differences between the equating functions based on the repeater group data and the first-time examinee group data to see whether there was a significant difference in equating outcomes between the two subgroups. These two sets of comparison outcomes should be fairly consistent.

Equating Models

In deriving equating functions for the total group and its two subgroups, we considered both the linear and non-linear equating models. Specifically, for each of the study equatings we produced equating functions based on the Tucker, chained linear, and smoothed chained equipercentile models. After a careful review and comparison of the various equating functions, we selected an equating conversion that best fit a particular group data. This way, the equating functions derived for the total group and its subgroups could be based on different equating models but the respective equating conversions would be optimal in meeting operational equating evaluation/selection criteria. While the selected equating conversions based on this approach would not be subject to bias due to the use of one single equating model, differences between equatings could be subject to model effects. Nevertheless, we considered the potential drawbacks of model effects less serious than the problems associated with applying just one equating model for all of the study groups.

For example, the best equating model for the total group might be the smoothed chained equipercentile equating but the model that best fit the first-time examinee group and/or the repeater group data could be linear, especially when the size of some subgroup(s) was small. If we only considered one equating model for all of the groups (or



subgroups), the adequacy of the equating functions might be compromised, and this effect might confound with the repeater effect that we aimed to study.

A Focus on Raw-to-Scale Equating

In this study, we chose to focus on equating that converts new-form raw scores to scaled scores used for score reporting (i.e., the raw-to-scale equating), because in practical equating situations raw-to-scale equating results are much more critical [than that for the raw(new)-to-raw(reference) equating] in terms of test consequences. In other words, it is more meaningful to investigate the raw-to-scale equating in the context of score fairness.

Technically, in equating research it may be more complex to study raw-to-scale equating because of the need to composite the raw(new)-to-raw(reference) and the reference-to-scale equating functions, which not only adds complexity to equating process but can also complicate the evaluation of equating outcomes. For instance, special consideration/treatment is needed for determining the scale score values for equated raw scores that go beyond the reference-form possible score range (i.e., when impossible scaled score conversions occur)³. And, in some cases the method (e.g., the linear interpolation approach) used for combining two equating functions may introduce bias to the final scaled scores. The reliance on the reference-to-scale equating function (of the raw-to-scale equating) in our study also represents a trade-off between equating practicality (i.e., utility) and equating precision. We will further discuss this trade-off in the discussions section.

Summarizing Equating Differences

We present detail of equating differences across various study groups (in scale score unit) by new-form raw score levels in graphs. The graphical presentations help to show the direction and magnitude of equating differences along the new-form score scale.

³ In this study, we decided to extend the reference-to-scale equating function to obtain scaled scores for the out-of-range equated raw scores but then truncate the scaled scores at the possible min/max (i.e., 20/80) on the score-reporting scale for study purposes. This approach worked because practically all the “imputed” scaled scores went beyond the possible min/max and ended up being truncated to the min/max values. The only and minor drawback was that a very small number of real (i.e., not imputed) scaled scores with values greater than the possible min/max also got truncated.



Using the set of equatability indices (also known as the score equity analysis, SEA, indices) developed for checking the subpopulation invariance properties of an equating function and for checking the equity of scores across subpopulations (Dorans & Holland, 2000; Dorans, 2004; von Davier, Holland, & Thayer, 2004; Yang, 2004), we also summarized the comparison outcomes between the total-group equating function and the equatings for its respective subgroups (i.e., the repeater and the first-time examinee groups). Used widely in a series of studies since 2000 (Liu, Cahn, & Dorans, 2006; von Davier & Wilson, 2008; Dorans, Liu, & Hammond, 2008; Liu & Holland, 2008; Yang & Gao, 2008; Yi, Harris, & Gao, 2008), the set of summary statistics will help to assess adequacy of the total-group equating function (and the first-time-examinee group equating function) and contrast the overall differences between the total-group and the first-time examinee group equating functions.

Specifically, the summary statistics we used include the root mean square difference (*RMSD*), the root expected square difference (*RES_{Dj}*), and the root expected mean square difference (*REMSD*). The *RMSD* summarizes the differences between the total and the subgroup linking functions across subgroups at various score levels, the *RES_{Dj}* evaluates the linking differences between each subgroup and the total group across score levels, and the *REMSD* is an overall measure of differences between the total and the subgroup linking functions across subgroups and score levels. The formulas for computing these statistics are presented below:

Let *P* be the population of examinees (for the new-form administration), with subpopulations *P_j* that partition *P* into two (i.e., *J*=2) mutually exclusive and exhaustive subpopulations, namely, the repeater and the first-time examinee groups. The *RMSD* can be computed as:

$$RMSD(x) = \sqrt{\sum_{j=1}^J w_j \left[e_{P_j}(x) - e_P(x) \right]^2},$$

where *x* is a raw score level on the new form, *e_P*(*x*) denotes the raw-to-scale equating function that places *x* on the reported score scale for the total population *P*, *e_{P_j}*(*x*) denotes the raw-to-scale function that places *x* on the reported score scale for the subpopulation



P_j , w_j is the proportion of P_j in P , and $\sum w_j=1$ (Dorans & Holland, 2000; Dorans, 2004; von Davier, Holland, & Thayer, 2004). Like P and P_j , w_j is defined in the context of the new-form administration.

As a weighted average of differences between a subpopulation linking function and the total group linking function, the $RESD_j$ can be calculated as follows:

$$RESD_j = \sqrt{E_P \left\{ \left[e_{P_j}(x) - e_P(x) \right]^2 \right\}} = \sqrt{\sum_{x=0}^Z w_{xP} \left\{ \left[e_{P_j}(x) - e_P(x) \right]^2 \right\}},$$

where j denotes a subpopulation, $E_P\{ \}$ denotes averaging over raw score levels weighted by the relative number of examinees at each score level in the total population P , Z is the maximum possible raw score, w_{xP} is $\frac{n_x}{n}$ in the total population P , and $\sum w_{xP}=1$. Note that n_x is the number of examinees at raw score level of x , and n is the total number of examinees (Yang, 2004). In addition, P , P_j and w_j are all defined in the context of new-form administration.

Summarizing the linking differences across score levels and subpopulations, the $REMSD$ can be calculated using the formula below (Dorans & Holland, 2000; Dorans, 2004; von Davier, Holland, & Thayer, 2004):

$$REMSD = \sqrt{\sum_{j=1}^J w_j E_P \left\{ \left[e_{P_j}(x) - e_P(x) \right]^2 \right\}}.$$

And, the above formula can be expanded as below (Yang, 2004; Yang & Gao, 2008):

$$REMSD = \sqrt{\sum_{j=1}^J w_j \sum_{x=0}^Z w_{xP} \left[e_{P_j}(x) - e_P(x) \right]^2} \text{ or } \sqrt{\sum_{x=0}^Z w_{xP} \sum_{j=1}^J w_j \left[e_{P_j}(x) - e_P(x) \right]^2}.$$

Other than the above summary statistics, we used statistics of simple differences to summarize scaled score differences between equatings in the repeater group, first-time examinee group and total group.



Score Difference that Matters

To determine whether the equating differences in scaled scores were of practical significance between various study groups, we compared the magnitude of the scaled score differences (and the statistics used to summarize these differences) to a criterion that represented the critical score difference that mattered (DTM) to the study exam. Specifically, the criterion for evaluating practical significance of scaled score differences was based on half a score point on the subscore scale of the study exam.

In addition to reporting the composite score, which is the sum of the weighted subscores for the four component measures, the study exam also reports subscore for each of the four measures on a 20 to 80 integer score scale. The four subscores are as important as the composite score to the examinees and the test users, because various graduate programs in different major fields may lay differential emphases on the four skills and require their applicants to meet different standards on these four measures. Although there may not be a consensus on what score difference would matter to the institutions that accept the scores on the study exam, it seems appropriate to say that in general a one-point difference on the subscore scale is a DTM to the examinees taking the study exam and to the test users, because graduate institutions often set a cutscore to screen their applicant pools and one score point difference on the subscore scale could translate to several points on the composite score scale for the study exam. Furthermore, because operationally half a score point on the subscore scale would be rounded to 1 for score reporting purpose, it seems more appropriate to define the DTM for this study as half a score point on the 20-to-80 subscore scale. From a practical perspective, we would consider an equating difference negligible if it is smaller than the DTM.

Simulations for Standard Error Estimation

To estimate standard errors of equatings (SEEs), standard errors of equating differences (SEEDs), and standard errors of subpopulation invariance measures (a.k.a. the equatability or SEA indices), we treated four of the smoothed (test, anchor) bivariate distributions (which were for the repeater and first-time examinee groups on the new and



reference forms⁴) that were used for the smoothed chained equipercentile equatings as population distributions and drew 500 random samples (with replacement) of the size of the original data from each of these distributions. We then generated the equating functions, scaled scores, scaled score differences, and subpopulation invariance measures for the 500 simulated samples, and used the standard deviations of the scaled scores, scaled score differences, and subpopulation invariance measures over the 500 samples to estimate the corresponding standard errors.

The standard error estimates of equating differences (i.e., the SEEDs) served as a criterion for evaluating the statistical significance of equating differences in scaled scores between study subgroups, and the standard error estimates of the subpopulation invariance measures were used to evaluate the statistical significance of the scaled score differences between the total group and its subgroups (Moses, 2006). Given the relatively small study sample size, especially the size of the repeater groups, it was crucial to evaluate the statistical significance of the equating differences (on the scale of reported scores) to determine whether study findings were subject to sampling errors.

In summary, we could justify the use of (only) the first-time-examinee group data for equating if the repeater effects on score equating were significant (i.e., the equating differences between various study groups were statistically and/or practically significant). If the repeater effects were not significant, it might not be necessary to exclude the repeaters from the equating samples. By excluding repeaters from equating when the repeater effects were not significant, one may inadvertently lower the equating precision due to the reduction in equating sample size and the potential alteration of equating sample representation.

Results

We first present the identification/verification outcomes for the self-reported repeater data to set the grounding for this study. Then, we present various analysis outcomes that describe the general trends of repeater performance, followed by the

⁴ Simulated data for the repeater and first-timer groups were combined and used as the basis for estimating the standard errors of various statistics of interest for the total group.



results of repeater effects on score equating. An understanding of the trends in repeater performance will aid the interpretation of the results for the repeater effects on equating by putting the equating results in proper perspective.

Identification/Verification of Repeater Status

Overall, we found a nearly 88% match (72% non-repeaters and 16% repeaters) between the repeater groups identified by the voluntary self-reporting survey approach and the approach based on matching the empirical examinee data across administrations. While the self-reported approach identified the other 11% or so examinees as repeaters the empirical approach did not agree, which was likely a miss by the empirical approach due to the imperfect matching of examinees' records across administrations. In addition, the empirical approach only picked up about 0.5% of the examinees as repeaters while the self-reporting approach indicated the opposite. From administration to administration, the actual percentage of match/mismatch between the self-reporting approach and the empirical approach varied. The empirical approach consistently yielded a lower repeater rate than the self-reporting approach across administrations; the differences could be as large as 16% for one administration, which did not seem realistic at all. In short, the empirical approach was much more likely to miss real repeaters than picking up those not identified by the self-reporting approach (i.e., those examinees that concealed their repeater identity in the voluntary repeater survey).

The disagreement between the repeater identification outcomes based on the empirical approach and those based on the self-reporting approach was probably due to the lack of reliable and effective matching variables for merging examinee records across exam administrations. If there were a more effective way to empirically identify the real repeaters, we could even avoid using the voluntary, self-reported repeater information and instead use the empirically identified repeater data for our study. However, none of the available matching variables, or any of the combinations of these variables, seemed to work well enough to produce trustworthy empirical repeater data that was more reliable than the self-reported repeater data. In other words, the empirical identification approach was deemed not feasible. Therefore, we decided to use the self-reported repeater data for our analyses in this study, mainly to avoid the under-identification of real repeaters. In



general, the self-reported repeater data looked reasonably sound. Although it was not perfect, it was the best option we could have for this study.

General Trends of Repeater Performance

Across various study administrations, most of the examinees in the general, non-specific (i.e., not targeted at any number of re-takes, any specific time interval between test and retest, etc.) repeater group were 20 to 50 years old, with a concentration between 21 and 30 years of age. Based on the merged examinee data across administrations, we found that about 10% of the examinees repeated the exam only once, about 2.5% repeated twice, about 1% repeated three times, and less than 1% repeated more than three times. The actual repeater rates at different retake levels were likely to be higher than those reported above, because of the difficulty in effectively matching empirical examinee data across administrations, as explained previously. Despite the limitation, the empirical findings on the number of retakes still offered useful insights for studying the general repeater trend and patterns, especially when the self-reported repeater data did not provide such information at all (the repeater survey of the study exam was not designed to collect such information).

Test-Retest Correlations

For the overall, non-specific repeater group (N=6,256), the test-retest correlation coefficient was 0.74 for Verbal and 0.72 for Quantitative, which were based on the repeaters' scale scores from the two most recent testing occasions. The magnitude of the positive correlation coefficients looked reasonable and was typical of the test-retest correlations for exams measuring similar constructs. The test-retest correlation result suggested a somewhat strong relationship between the test and retest scores for the overall repeater group. Nevertheless, extra cares are needed when interpreting or generalizing this result. Because repeaters are usually self-selected (i.e., not randomly representative of the total examinee group), study outcomes based on the repeater data are subject to range restriction problems, which may not be generalized to the entire examinee population.

For Verbal and Quantitative measures, respectively, Table 3 presents the test-retest correlation coefficients for various repeater subgroups that differed in the time interval between testing occasions. Overall, for both of the two measures the magnitude



of the test-retest correlation coefficient (r) seemed to be independent of the time interval (t , in years) between testing occasions. However, for Verbal the magnitude of r increased while t increased until t reached 4 (years), and r decreased when t increased from 4 to 7. These opposite trends for different time frames apparently cancelled each other out and contributed to the overall impression that r was independent of t for Verbal.

Anyway, the above findings are not in agreement with the common belief that r would decrease while t increases because the practice and/or the recency effect is likely to diminish when time goes by. Perhaps in future we could look into repeaters' academic status when they take and retake the same exam to see whether the differences in examinees' academic status could help to explain the current study findings.

Table 3. Test-Retest Correlation for the Repeater Group

Test Measure	Time Interval, t , between Testing Occasions (Years)	n	Test-Retest Correlation Coefficient (r)
Verbal	$0 < t < 1$	4,498	0.72
	$1 \leq t < 2$	850	0.74
	$2 \leq t < 3$	402	0.77
	$3 \leq t < 4$	253	0.80
	$4 \leq t < 5$	168	0.78
	$5 \leq t < 6$	78	0.73
	$6 \leq t < 7$	7	-
<i>Verbal Overall ($0 < t < 7$)</i>		6,256	0.74
Quantitative	$0 < t < 1$	4,498	0.71
	$1 \leq t < 2$	850	0.69
	$2 \leq t < 3$	402	0.70
	$3 \leq t < 4$	253	0.76
	$4 \leq t < 5$	168	0.72
	$5 \leq t < 6$	78	0.73
	$6 \leq t < 7$	7	-
<i>Quantitative Overall ($0 < t < 7$)</i>		6,256	0.72

Note. The scale scores of repeaters from the two most recent testing occasions were used for this analysis. Tests taken in the prior or later occasion might not be the same for various examinees, who came from a wide range of test administrations.



Repeater Score Gain/Loss

An inspection of the general scale score gain/loss for the overall, non-specific repeater group revealed that...

- Overall, for the Verbal measure close to 59% of the repeaters improved their scores after retaking the study exam, more than 35% had score decreases, and about 6% saw no score change;
- Findings for the Quantitative measure were very similar to those for Verbal with slight differences;
- By repeating the study exam, on average the overall repeater group improved their scale scores by only about 2.2 points on either Verbal or Quantitative; however, the variability of scale score change was very large (the standard deviation was about 7.2 for either one of the measures), suggesting that large score gains and losses were canceled out during the summing and averaging process, which was especially true for the large study data spanning a large number of administrations.

The conditional repeater score gain/loss distributions provided an in-depth look at the trend of repeater performance. They allow comparisons of repeater score gain/loss across various repeater subgroups differing in their performance on the previous testing occasion. Table 4 presents the conditional distributions for the overall, non-specific repeater group, and Table 5 shows the conditional distributions for the much smaller group of repeaters taking the most recent study test.



Table 4. Conditional Scale Score Gain/Loss for the Overall, Non-specific Repeater Group

Test Measure	Scale Score on the Prior Test	% of Examinees with Different Degrees of Score Gain/Loss (Score on the Later Test - Score on the Prior Test)									Average Scale Score on the Later Test	Average Scale Score Gain/Loss
		Below -15	-11 to -15	-6 to -10	-1 to -5	No gain/loss	+1 to +5	+6 to +10	+11 to +15	Above +15		
Verbal	71-80	14	14	14	43	0	14	0	0	0	65	-8
	61-70	4	5	21	37	4	22	7	0	0	61	-3
	51-60	2	4	14	29	7	27	13	3	1	54	0
	41-50	1	3	9	21	6	29	20	10	3	48	2
	31-40	0	2	7	18	4	29	22	13	5	40	4
	20-30	0	0	4	12	6	25	25	19	10	33	7
Quantitative	71-80	8	4	19	23	10	26	10	0	0	71	-3
	61-70	2	7	14	28	6	24	14	5	0	63	-1
	51-60	1	4	13	24	6	28	16	6	2	56	1
	41-50	0	2	10	21	6	26	21	9	4	48	3
	31-40	0	0	3	16	5	30	25	14	7	42	5
	20-30	0	0	0	6	3	20	30	26	14	38	9

Note. The scale scores of repeaters from the two most recent testing occasions were used in this analysis. In this table, the test taken earlier was referred to as the "Prior Test" and the test taken later was referred to as the "Later Test". Tests taken by different repeaters in the prior or later occasion might not be the same test because examinees in the overall, non-specific repeater group came from a wide range of administrations.

Table 5. Conditional Scale Score Gain/Loss for the Repeat Group Taking the Most Recent Study Test

Test Measure	Scale Score on the Prior Test	% of Examinees with Different Degrees of Score Gain/Loss (Score on the Later Test - Score on the Prior Test)									Average Scale Score on the Later Test	Average Scale Score Gain/Loss
		Below -15	-11 to -15	-6 to -10	-1 to -5	No gain/loss	+1 to +5	+6 to +10	+11 to +15	Above +15		
Verbal	71-80	-	-	-	-	-	-	-	-	-	-	-
	61-70	8	8	0	31	8	31	15	0	0	62	-1
	51-60	3	0	18	29	6	25	8	8	3	55	0
	41-50	2	2	7	20	11	23	24	10	2	48	2
	31-40	0	1	1	17	5	33	21	16	4	41	5
	20-30	0	0	8	13	0	8	13	50	8	35	8
Quantitative	71-80	0	20	20	40	20	0	0	0	0	70	-6
	61-70	4	16	20	24	0	28	8	0	0	60	-4
	51-60	3	4	20	27	4	26	12	3	0	54	-1
	41-50	1	4	17	30	6	23	14	1	4	46	0
	31-40	0	0	4	16	12	33	22	12	2	41	4
	20-30	0	0	0	0	0	40	60	0	0	35	6

Note. The scale scores of repeaters from the two most recent testing occasions were used in this analysis. In this table, the most recent study test was referred to as the "Later Test" and the test taken earlier was referred to as the "Prior Test". Although all of the examinees in this analysis took the same "Later Test" (which was the most recent test investigated in this study), their "Prior Test" might vary.

The last column in Table 4 shows the average scale score changes for various repeater subgroups. The average change ranged from -8 to 7 across various subgroups for Verbal and from -3 to 9 for Quantitative. These results were quite different from the



average scale score change of 2.2 for the overall repeater group. The average scale score changes for various repeater subgroups also indicate that overall the higher the repeaters' scores on the prior test, the lower their score gains on the later test for both Verbal and Quantitative. There were even negative score gains (i.e., score loss) for the repeaters scoring above 60 on the reporting scale for both measures. This result looked reasonable and was not surprising because of the ceiling and regression to the mean effects. The average scale scores on the later test, as reported in the second last column of the table, looked consistent to this result. This implies that the high-performing examinees may not increase their scale scores by retesting (the retest scores could be even lower than before), while the low-performing examinees could improve their scale scores by a substantial number of points by retesting.

The numbers in the nine columns in the middle of Table 4 are the percentages of examinees with different degrees of score gain/loss (on the later test), conditioned on the examinees' performance on the prior test. The conditional score gain/loss results (especially those presented in the shaded three columns at the center of the table) further indicated that in general the majority of the examinees (i.e., over 50%) in various repeater subgroups had a score change (either increase or decrease) that was equal to or less than five points, regardless of their performance on the prior test. An exception was the worst-performing examinees on the prior test, for which the majority gained more than five points after retaking the exam. And, generally, the lower the scores on the prior test, the larger proportions of repeaters with score changes that were greater than 5 points. For examinees at the lower score levels on the prior test, most of the score changes on the later test were positive. We also found the conditional distribution results based the most recent study test data to be fairly consistent with the results for the overall, cross-administration data.

Performance of Repeaters vs. First-Time Examinees on New Forms

Analyses of repeater performance based on reported scale scores were subject to potential equating bias that we set out to investigate in this study, because the underlying assumption of these analyses was that the practice of excluding repeaters from the equating samples would have no negative effect on the comparability of scale scores across administrations. Therefore, we also analyzed repeater performance by using the



raw (i.e. not equated) scores of various examinee groups on the same test form, which were not under the influence of subsequent equating practices.

Figures 1 to 5 present the observed relative frequency distributions of the repeater group, the first-time examinee group, and the total examinee group on the raw total-score scale for the five new forms, respectively. In each of the figures, there are two distribution plots—the plot on the left shows the percentage of examinees in a particular study group at each test score level for each of the three study groups (i.e., the repeaters, the first-timers, and the total group); and, the plot on the right shows the percentages of examinees in the total examinee group for various study groups. While the plot on the right reflects the proportions of the repeaters and the first-time examinees in the total examinee group, the plot on the left makes it easier to compare the shape and location of the score distributions for the various study groups that differed in size (by expressing the frequency distributions of various groups on a comparable percent scale).

Figure 1. Observed relative frequency distributions for Form A-V.

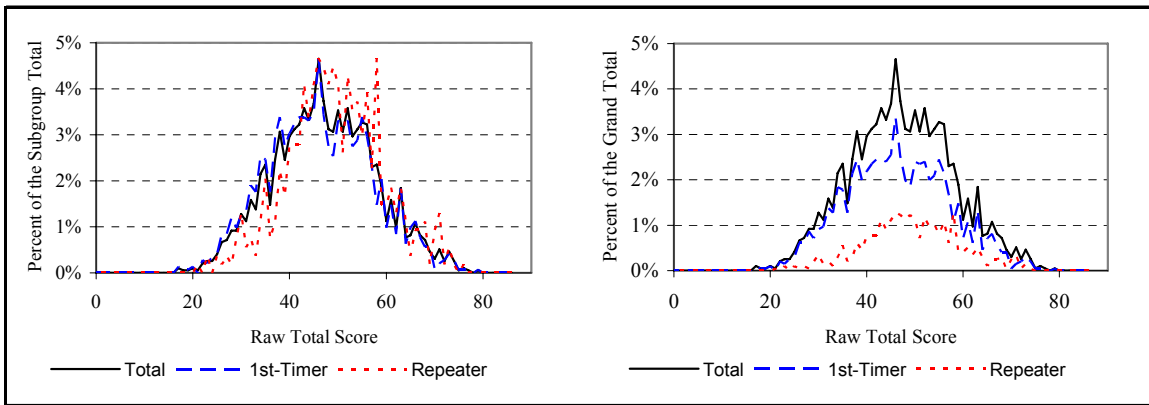


Figure 2. Observed relative frequency distributions for Form A-Q.

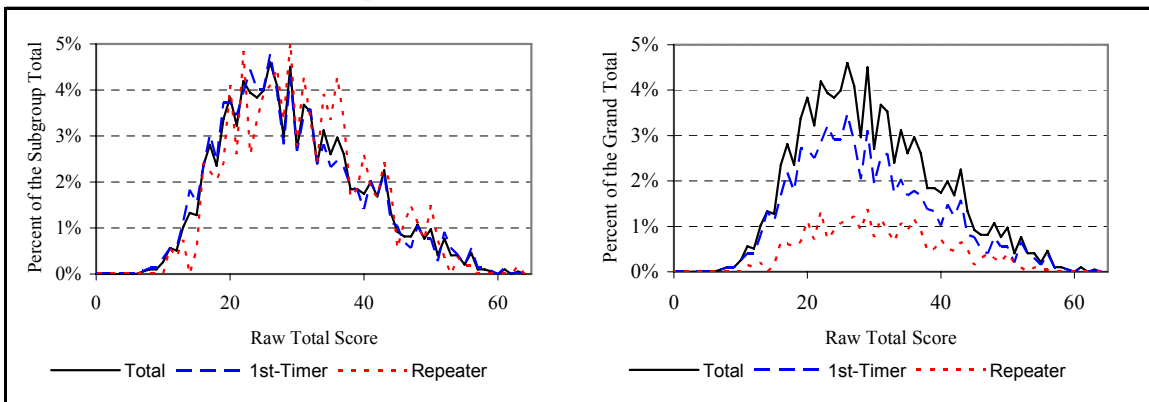




Figure 3. Observed relative frequency distributions for Form B-V.

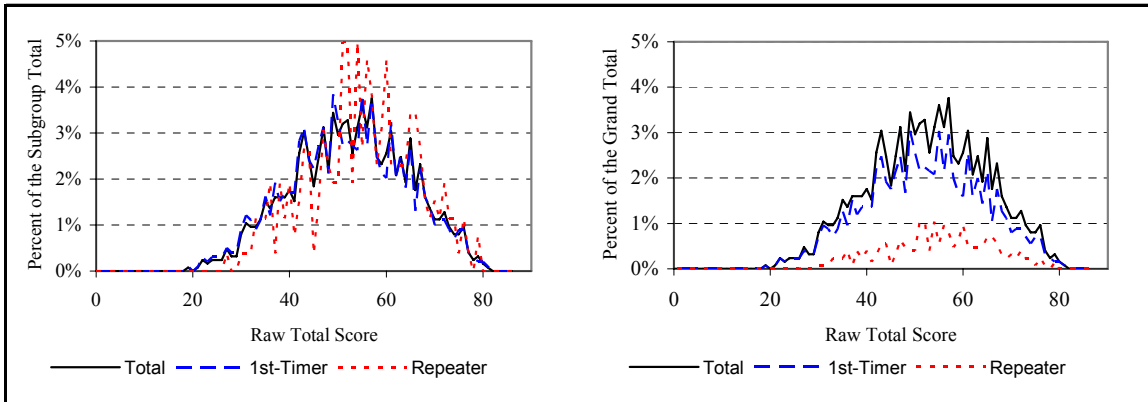


Figure 4. Observed relative frequency distributions for Form B-Q.

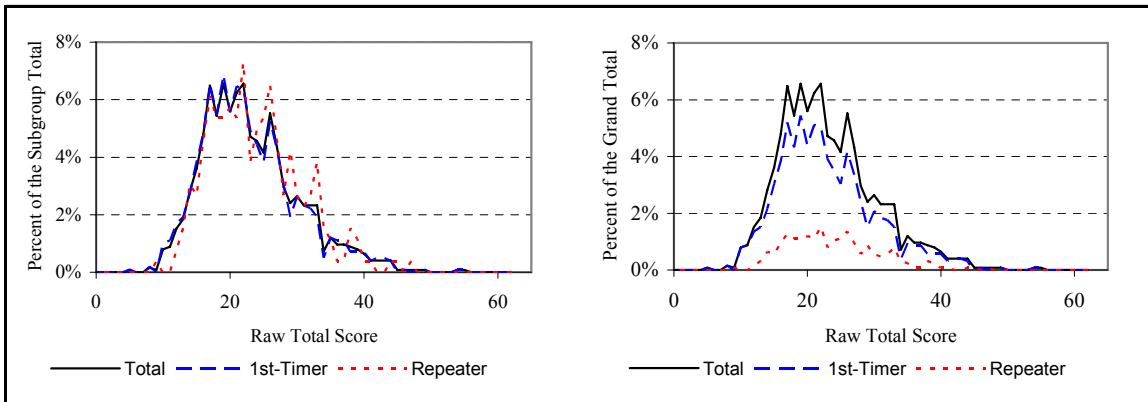
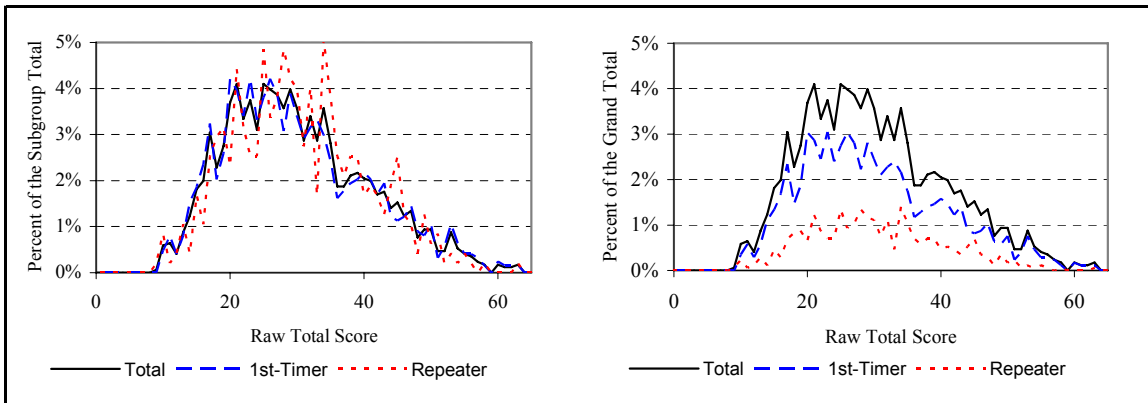


Figure 5. Observed relative frequency distributions for Form C-Q.



Overall, the figures show that the repeater score distributions looked quite similar to the distributions for the first-time examinees, despite their more distinctive ups and downs in frequencies. Nevertheless, the location of the distributions for repeaters was slightly to the right of that for the first-time examinees, especially at the lower end of the



raw score scale (although the differences on the plots do not look very pronounced). This suggests that on average the repeaters might have performed better than the first-time examinees on the study exam.

As indicated by the group means presented in Tables 1 and 2, the repeater group consistently performed better than the first-time examinee group on the total test (and on the anchor test) across various study forms, except for on two of the reference forms for the Quantitative measure (namely, RB-Q and RC-Q). Therefore, we conducted a two-sample Z test for each of the study forms to determine whether the repeater group had performed significantly different from the first-time examinee group. Table 6 shows the Z test results across study forms (for both new and reference forms).



Table 6. Significance of Group Mean Differences between 1st-Timers & Repeaters

Test Form		Examinee Group	n	Test Score		Z	p
				Mean	SD		
New	A-V	1st-timer	1,419	46.38	10.79	5.91	<.0001
		Repeater	537	49.42	9.90		
	A-Q	1st-timer	1,419	29.08	10.20	3.16	0.0016
		Repeater	537	30.63	9.47		
	B-V	1st-timer	989	52.08	12.19	3.60	0.0003
Repeater		261	54.90	10.98			
B-Q	1st-timer	989	22.92	7.26	1.90	0.0574	
	Repeater	261	23.83	6.79			
C-Q	1st-timer	1,234	29.97	10.77	0.60	0.5485	
	Repeater	474	30.30	9.89			
Reference	RA-V	1st-timer	1,239	49.22	11.01	4.70	<.0001
		Repeater	653	51.62	10.30		
	RA-Q	1st-timer	1,178	30.79	11.23	0.91	0.3628
		Repeater	453	31.31	9.93		
	RB-V	1st-timer	1,081	45.28	11.70	3.07	0.0021
Repeater		312	47.53	11.31			
RB-Q	1st-timer	1,081	24.48	7.21	-0.44	0.6599	
	Repeater	312	24.29	6.55			
RC-Q	1st-timer	1,753	38.56	11.50	-0.70	0.4839	
	Repeater	554	38.18	10.93			

Note. The null hypothesis was: $\mu_{repeater} - \mu_{first-timer} = 0$ for the two-tailed Z test.

In general, the Z-test results indicate that the mean scores of the repeater group and the first-time examinee group were significantly different on half of the 10 study forms. The repeater group performed significantly better on three of the new forms and on two of the reference forms. On the two reference forms that the repeater group scored slightly lower than the first-time examinee group, the differences were rather small and not statistically significant. Overall, these findings suggest that the repeaters are very likely to be more able than the first-time examinees on the study exam (at least, the repeaters are as able as the first-time examinees).



Equating Outcomes for Various Study Forms

With the above findings on quality of the repeater data and general trends of repeater performance in mind, we considered equating results for various study forms. For each of the equating functions needed (for the total group, the first-time examinee group, or the repeater group), we compared various equating results (based on the Tucker, chained linear, and smoothed chained equipercentile models) and selected an equating function that best fit a particular group data. The selection criteria were consistent with the criteria used operationally for evaluating equating outcomes for the study exam. Table 7 summarizes the equating model selection outcomes for each of the study groups on various forms.

Table 7. Selected Equating Model for Each Examinee Group on Each Form

New Form	Reference Form	Examinee Group	Equating Model Chosen
A-V	RA-V	1st-timer Repeater Total	Chained Linear Chained Linear Chained Linear
A-Q	RA-Q	1st-timer Repeater Total	Smoothed Chained Equipercentile Smoothed Chained Equipercentile Smoothed Chained Equipercentile
B-V	RB-V	1st-timer Repeater Total	Smoothed Chained Equipercentile Chained Linear Smoothed Chained Equipercentile
B-Q	RB-Q	1st-timer Repeater Total	Smoothed Chained Equipercentile Chained Linear Smoothed Chained Equipercentile
C-Q	RC-Q	1st-timer Repeater Total	Smoothed Chained Equipercentile Smoothed Chained Equipercentile Smoothed Chained Equipercentile

For each of the following new forms—A-V, A-Q, and C-Q, the same equating model was chosen for all of the three study groups. For either B-V or B-Q, the same equating model was chosen for the total and the first-time examinee groups, but a



different model was chosen for the repeater group (primarily because the smoothed chained equipercentile model produced unsatisfactory results and the sparse data of the repeater group could not support the use of the chained equipercentile model).

Repeater Effects on Score Equating

Evaluation outcomes for the repeater effects on score equating are presented in this section. For various study equatings, we tabulated results of the average subpopulation invariance (i.e., the $RESD_j$ and $REMSD$) for an overview. To provide more detailed information on how the scaled scores resulting from equatings in different study groups/subgroups differed, we graphed the scale-score differences between equatings and the $RMSD$ outcomes with a band of ± 2 standard errors (for evaluating the statistical significance) and a band for the DTMs (for evaluating the practical significance).

RESD_j & REMSD Results

For each of the five study equatings/administrations, Table 8 presents the $RESD_j$ and $REMSD$ results for evaluating the invariance of the total-group scaled scores with respect to the repeater and first-time examinee subgroups. The results in Table 8 are fairly consistent reflections of the characteristics of the invariance measures and the test data. The relatively small values and standard errors of the $RESD_j$ s for the first-time examinees were largely attributable to the large sizes of the first-time examinee subgroups across forms/administrations. The repeater subgroups were smaller and more distinct from the total group than the first-time examinee subgroup, so the values and standard errors of the repeaters' $RESD_j$ s were considerably larger than those for the first-time examinees' $RESD_j$ s. The $REMSD$ s that summarize the squared deviations of the repeaters' and first-time examinees' equating outcomes from the total group's equating outcomes had values that are in-between those of the repeaters' and first-time examinees' $RESD_j$ s.



Table 8. $RESD_j$ and $REMSD$ Results (with ± 2 Standard Errors in Parentheses)

New Form/ Reference Form	$RESD_j$		$REMSD$
	Repeaters	First-time examinees	
A-V/RA-V	0.2800 (± 0.4182)	0.1298 (± 0.1728)	0.1919 (± 0.2716)
A-Q/RA-Q	0.3995 (± 0.3348)	0.1220 (± 0.1039)	0.2793 (± 0.2217)
B-V/RB-V	0.7349 (± 0.5691)	0.1019 (± 0.1895)	0.3422 (± 0.2879)
B-Q/RB-Q	0.7283 (± 0.6358)	0.2174 (± 0.1913)	0.3928 (± 0.3269)
C-Q/RC-Q	0.3474 (± 0.4057)	0.0894 (± 0.1258)	0.2009 (± 0.2432)

Note. The range of practically significant subpopulation dependence is +0.5 and above, since the DTM for various study forms is 0.5 scale point.

While most of the $RESD_j$ s and $REMSD$ s in Table 8 did not exceed the practical significance criterion of +0.5 (the DTM for various study forms), two of the $RESD_j$ s for the repeaters (on Forms B-V and B-Q, respectively) were larger than 0.5, suggesting a practically significant equating difference in scaled scores between the total group and repeater subgroup. Overall, from a practical perspective the scaled scores based on equatings in the total group looked pretty invariant across subpopulations for various study forms, except that on two particular forms (one for Verbal and one for Quantitative) there might be a subpopulation dependence problem due to the equating differences in scaled scores between the total group and the repeater group.

In contrast to the results of practical significance, several more of the $RESD_j$ s and $REMSD$ s in Table 8 were statistically significant from zero (i.e., greater than +2 standard errors). Statistically significant $RESD_j$ s occurred for the repeaters on new forms A-Q, B-V and B-Q. Statistically significant $RESD_j$ s occurred for the first-time examinees on



forms A-Q and B-Q. And, statistically significant *REMSDs* occurred for A-Q, B-V and B-Q consequently.

Equating Differences in Scaled Scores and RMSD Results

To evaluate the invariance of scaled scores at individual new-form raw score levels, Figures 6-25 present the scaled score differences between equatings for the various study groups (i.e., the repeater group, first-time examinee group and total group), as well as the RMSD results, for the five study equatings/administrations. These figures provide more detail than the *RES_Ds* and *REMSDs* presented in Table 8, and Figures 6 to 20 further allow the positive and negative scaled score differences to be observed. Specifically, Figures 6-10 show the scaled score differences for the first-time examinee group and repeater group (First-time examinees – Repeaters). Figures 11-15 present the scaled score differences for the repeater group and total group (Repeaters – Total group). And, Figures 16-20 present the scaled score differences for the first-time examinee group and total group (First-time examinees – Total group). In addition, Figures 21-25 present the *RMSDs* for the five study equatings/administrations. The bands that indicate the DTMs and standard errors make it possible to evaluate the scaled score differences with respect to practical and statistical significance.



Figure 6
A-V/RA-V
Scaled Score Differences
(First-timers - Repeaters)

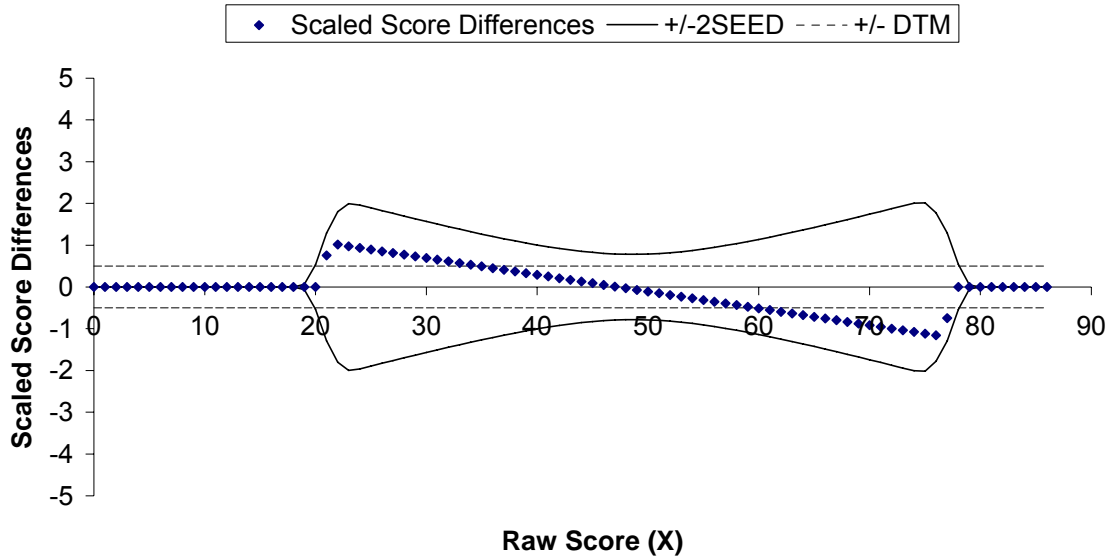


Figure 7
A-Q/RA-Q
Scaled Score Differences
(First-timers - Repeaters)

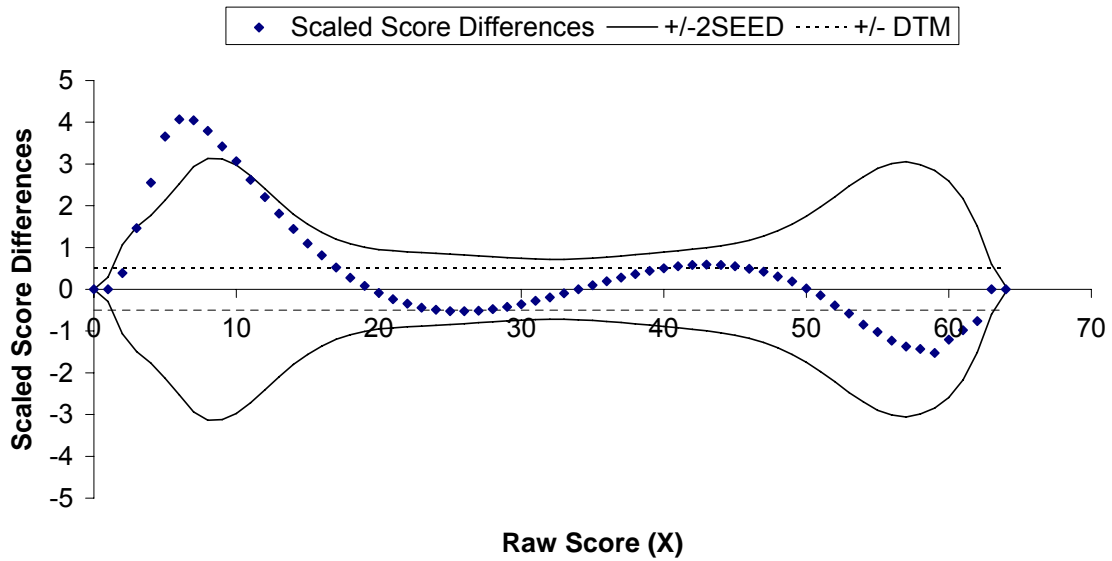




Figure 8
B-V/RB-V
Scaled Score Differences
(First-timers - Repeaters)

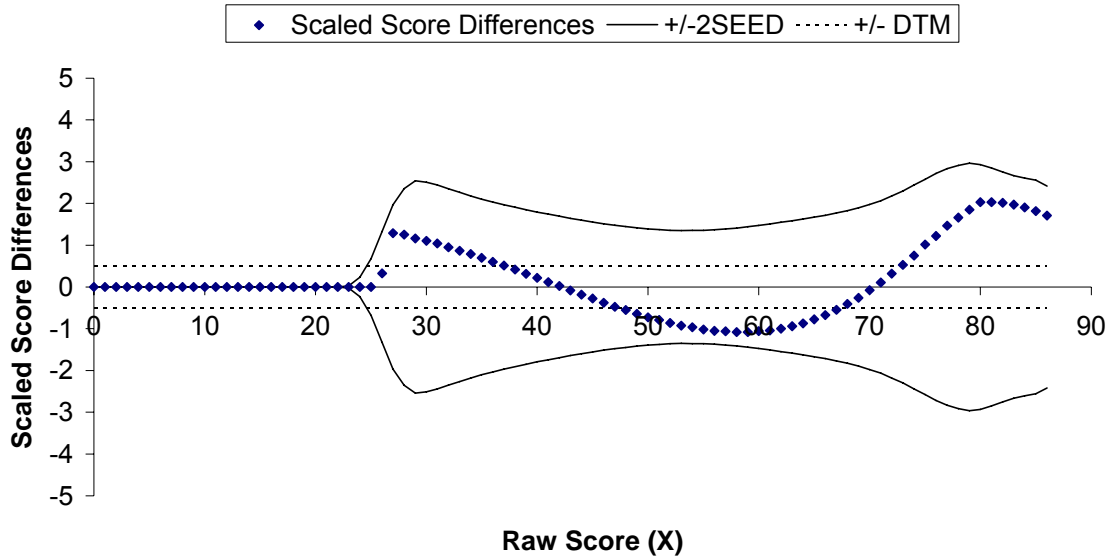


Figure 9
B-Q/RB-Q
Scaled Score Differences
(First-timers - Repeaters)

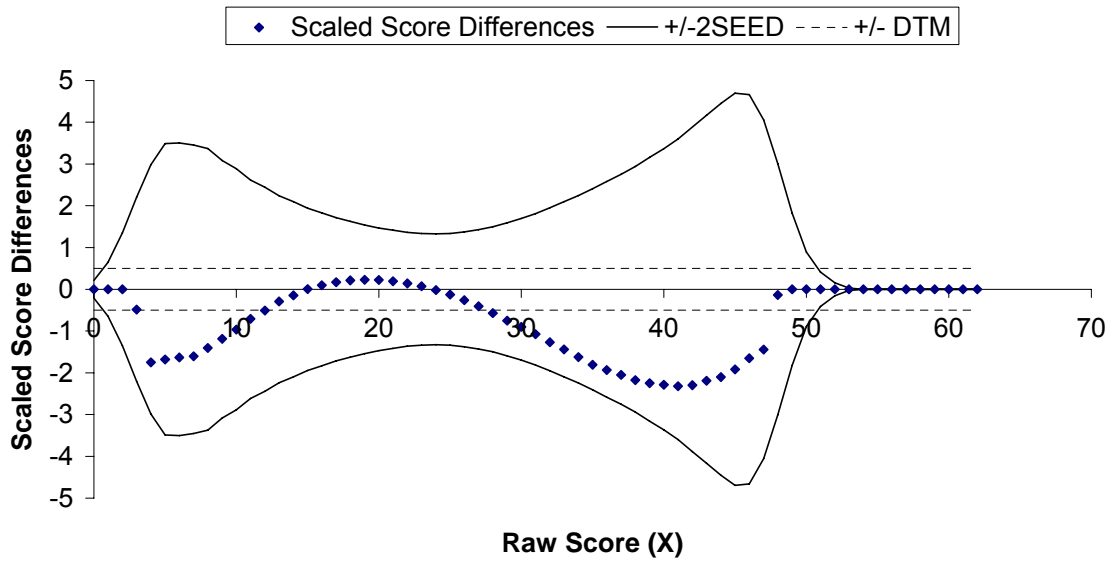




Figure 10
C-Q/RC-Q
Scaled Score Differences
(First-timers - Repeaters)

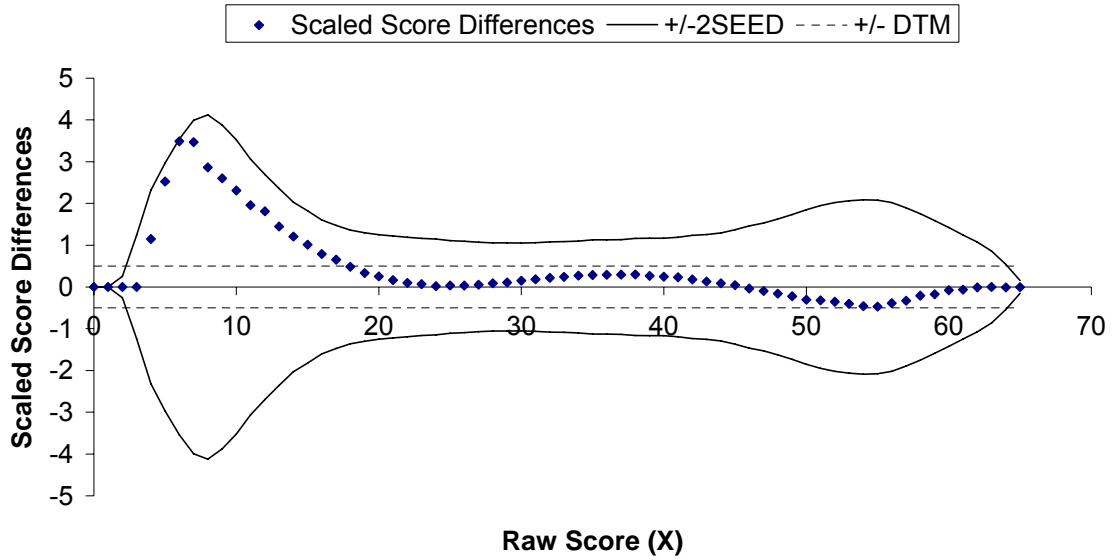


Figure 11
A-V/RA-V
Scaled Score Differences
(Repeaters - Total)

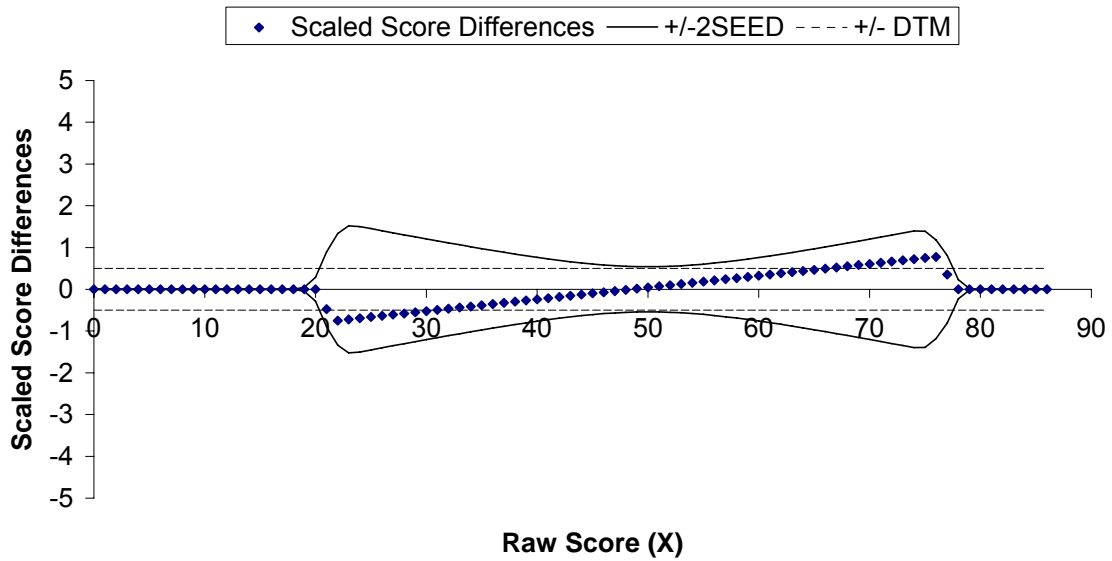




Figure 12
A-Q/RA-Q
Scaled Score Differences
(Repeaters - Total)

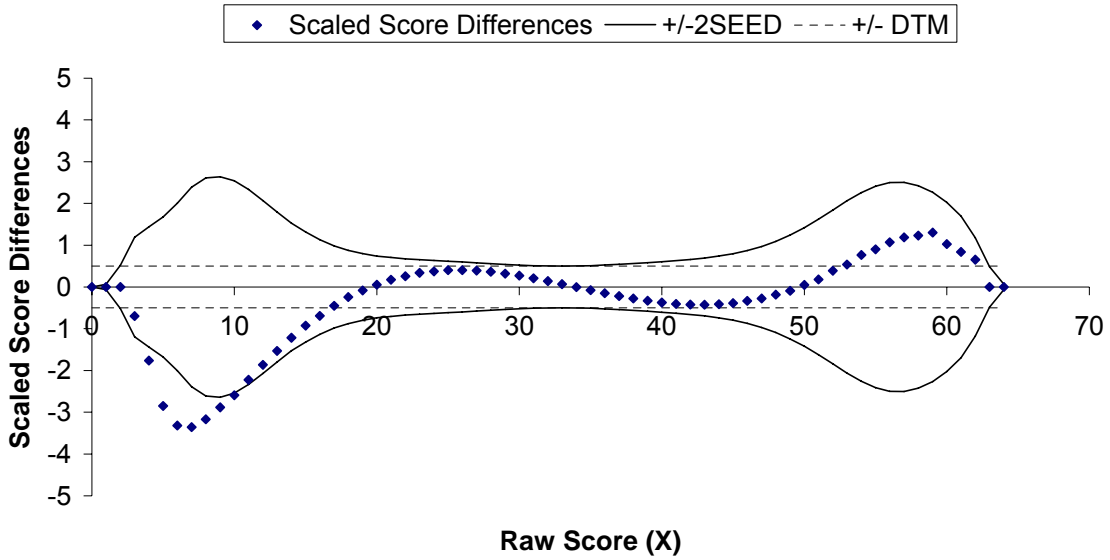


Figure 13
B-V/RB-V
Scaled Score Differences
(Repeaters - Total)

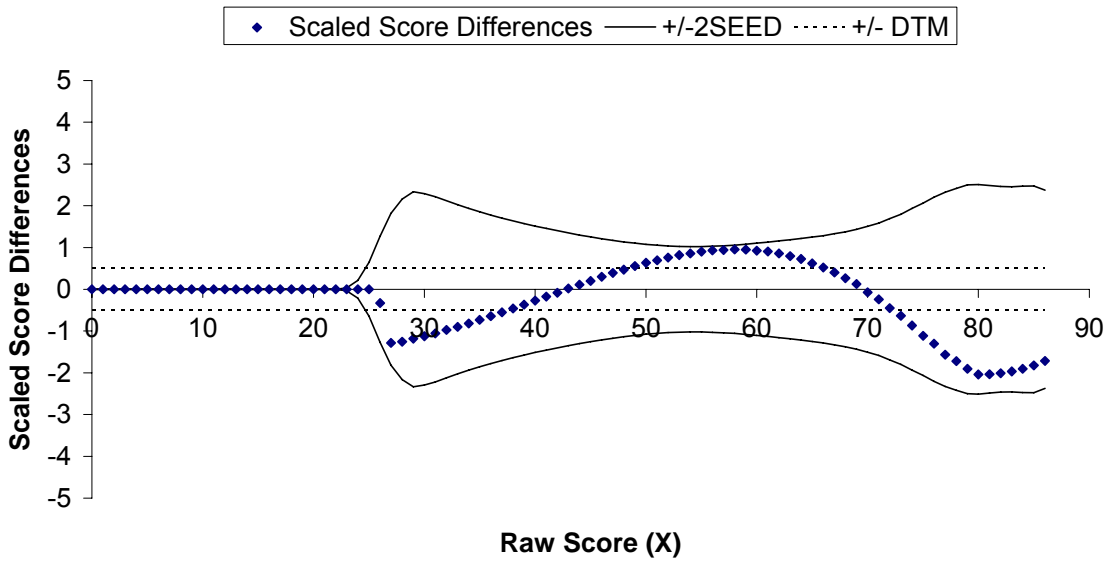




Figure 14
B-Q/RB-Q
Scaled Score Differences
(Repeaters - Total)

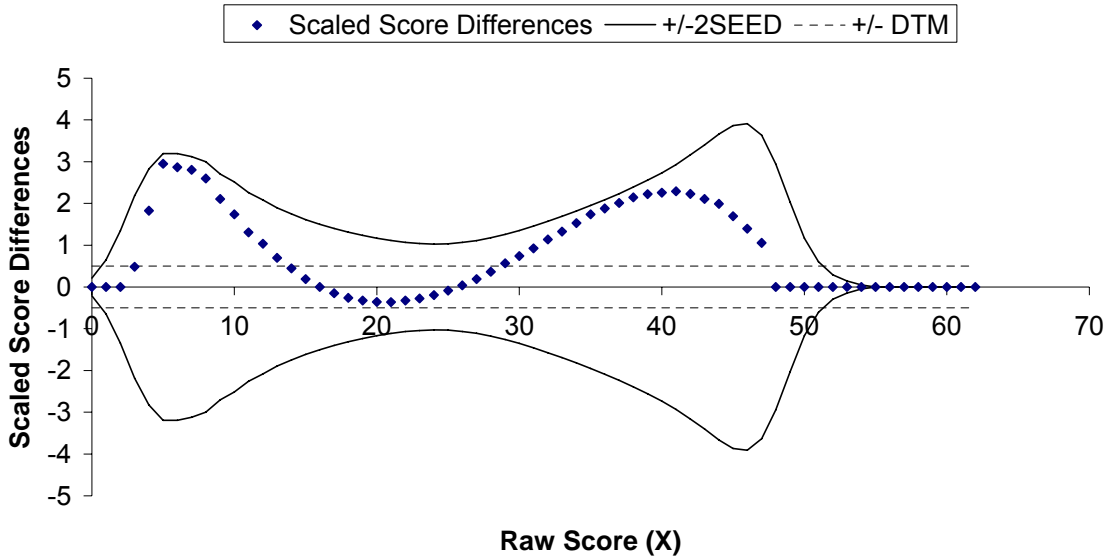


Figure 15
C-Q/RC-Q
Scaled Score Differences
(Repeaters - Total)

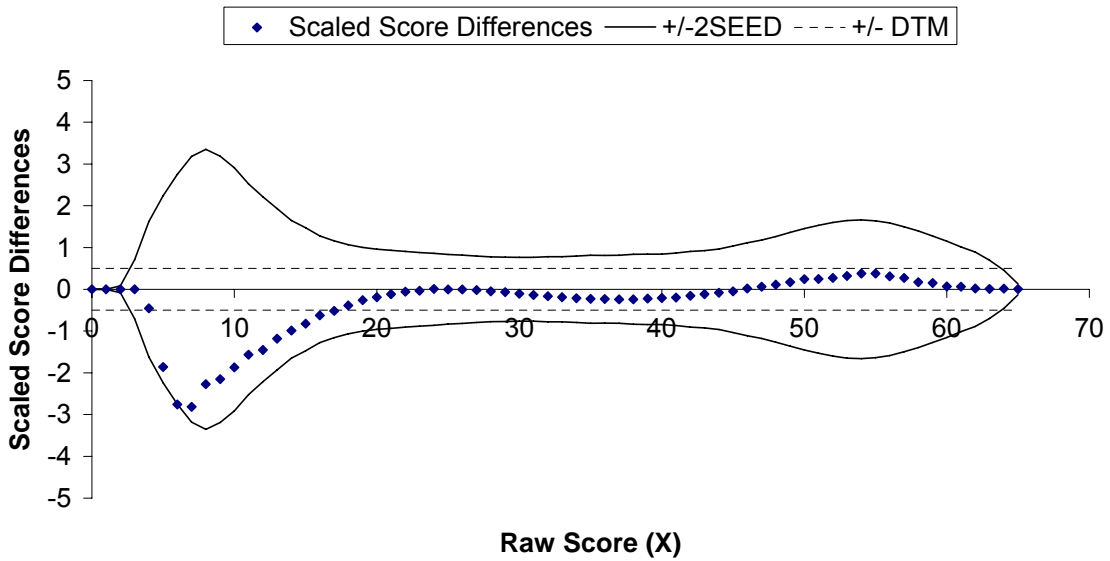




Figure 16
A-V/RA-V
Scaled Score Differences
(First-timers - Total)



Figure 17
A-Q/RA-Q
Scaled Score Differences
(First-timers - Total)

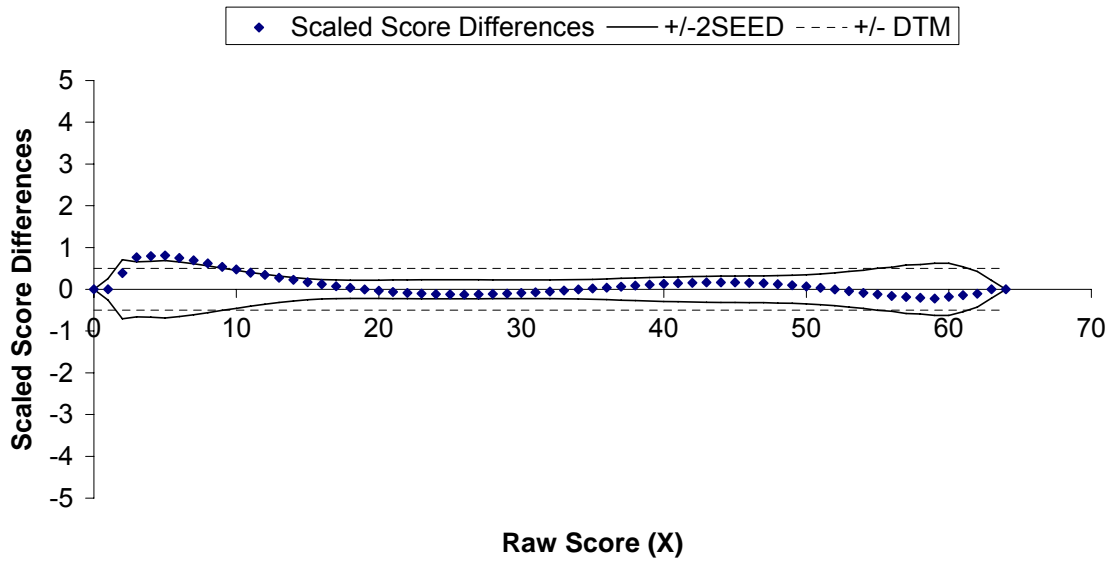




Figure 18
B-V/RB-V
Scaled Score Differences
(First-timers - Total)

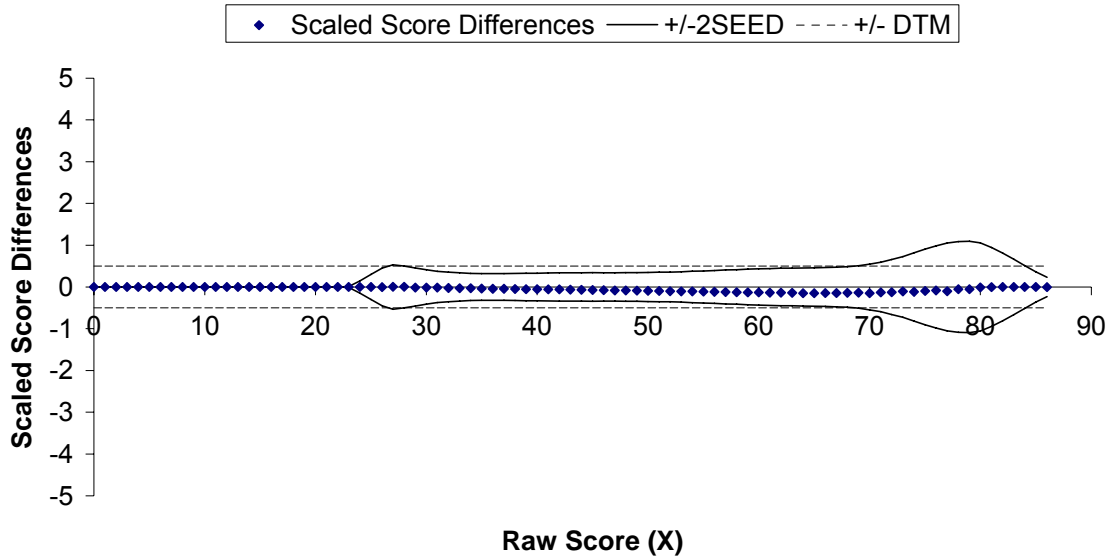


Figure 19
B-Q/RB-Q
Scaled Score Differences
(First-timers - Total)

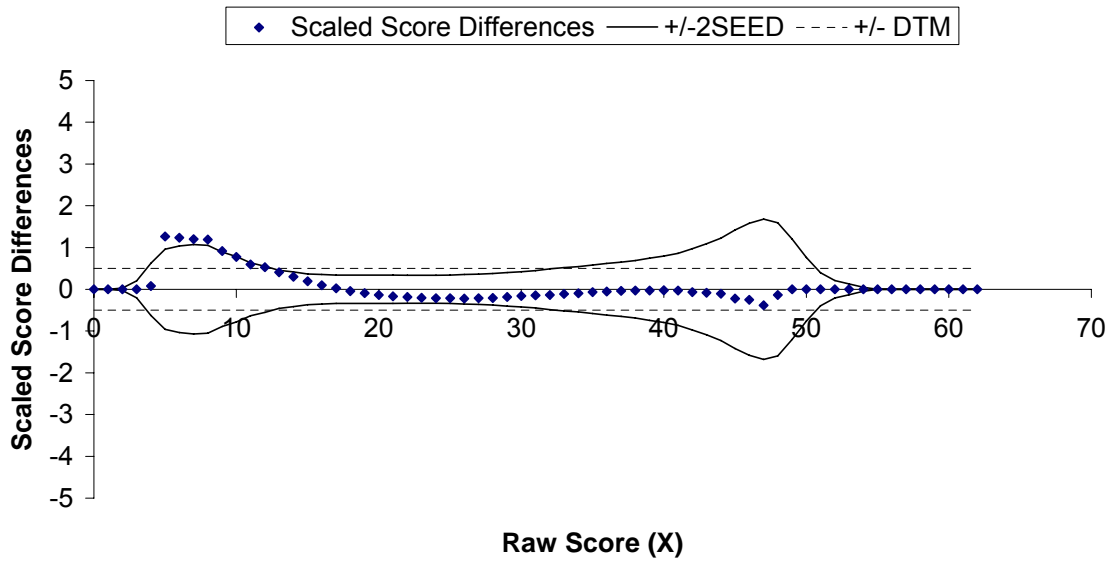




Figure 20
C-Q/RC-Q
Scaled Score Differences
(First-timers - Total)

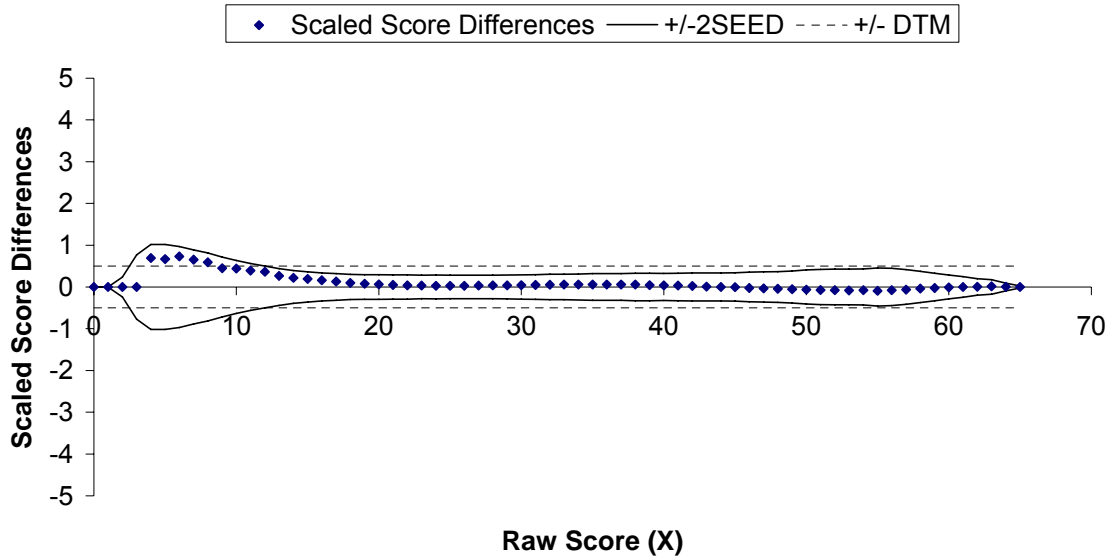


Figure 21
A-V/RA-V
RMSDs

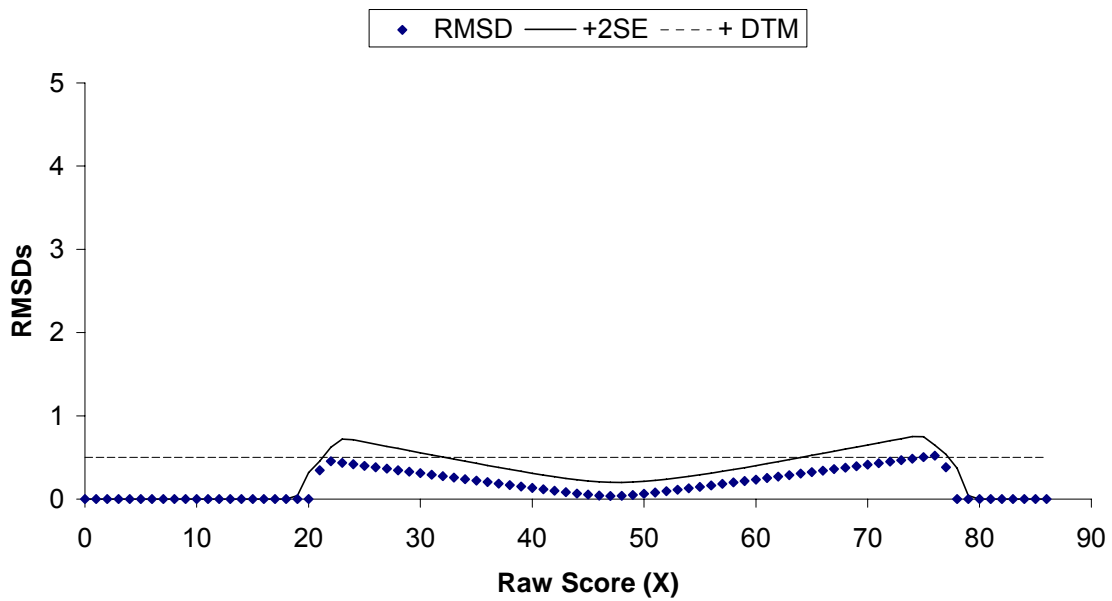




Figure 22
A-Q/RA-Q
RMSDs

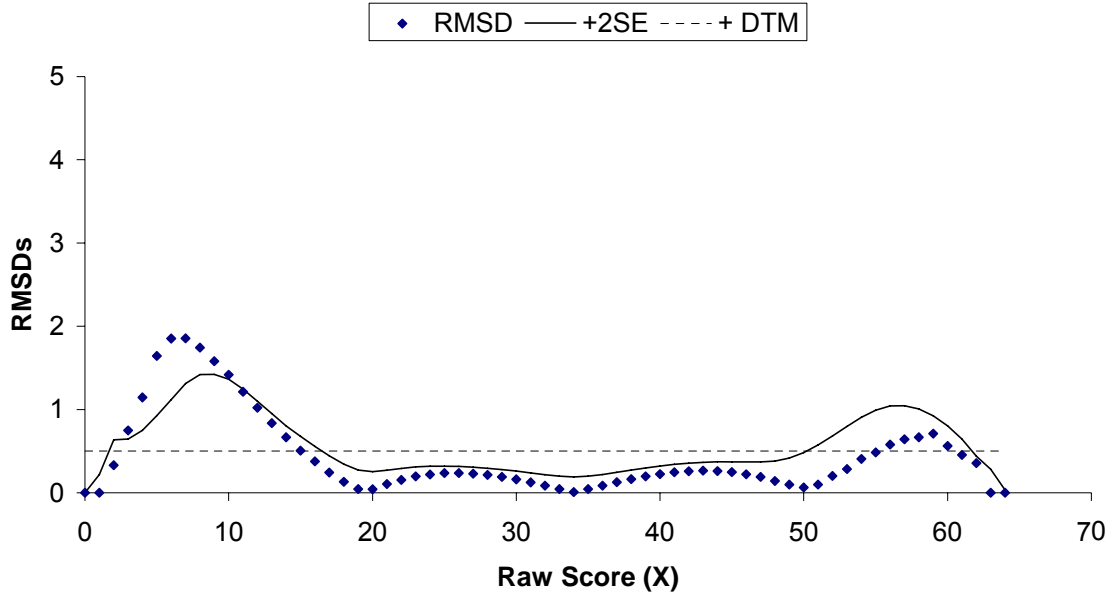


Figure 23
B-V/RB-V
RMSDs

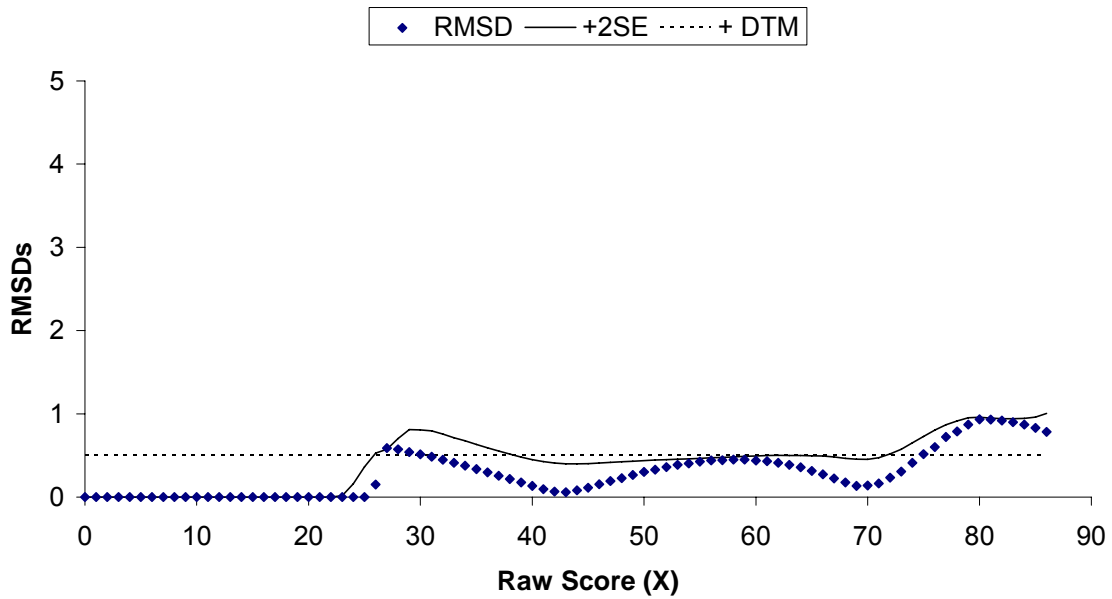




Figure 24
B-Q/RB-Q
RMSDs

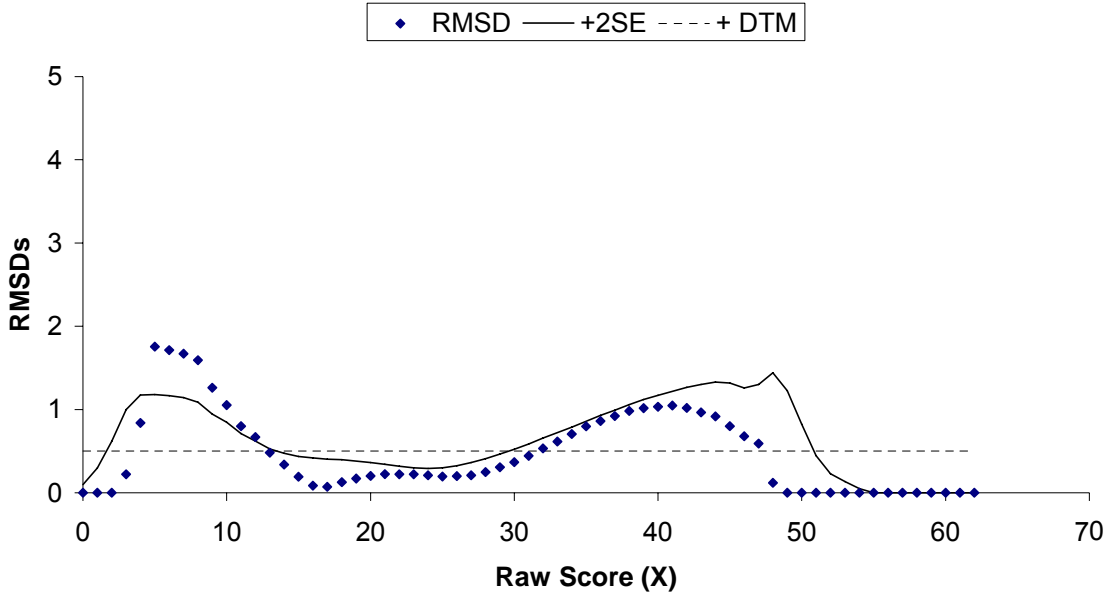
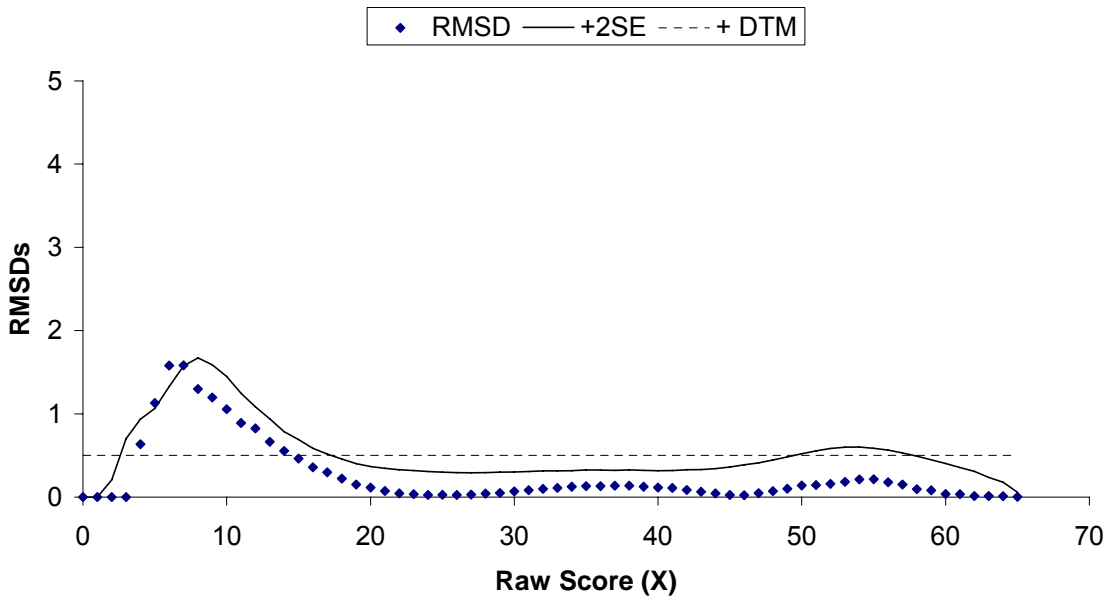


Figure 25
C-Q/RC-Q
RMSDs





➤ *Equating differences in scaled scores between first-timer and repeater groups*

Figures 6-10 compare the scaled score outcomes based on equatings in the repeater group and first-time examinee groups, which directly show how equatings in the two non-overlapping subgroups differed along the new-form raw score scale. Across the five study equatings/administrations, although Figures 6-10 show a large number of scaled score differences were of practical significance (especially for the A-V, A-Q, B-V, and B-Q administrations), almost none of the practically significant differences was statistically significant, which could be a result of the large standard error bands for evaluating statistical significance.

Despite that the scaled score differences shown in Figures 6-10 were often not statistically significant, directions of the scaled score differences generally supported the notion that the repeaters were at least as able as (and often more able than) the first-time examinees on the study exam. For example, Figure 6 shows that the equating in the repeater group yielded higher scaled scores than the equating in the first-timer group at the upper region of the new-form raw score scale, and Figure 9 shows that the equating in the repeater group consistently yielded higher scaled scores than the equating in the first-timer group along the raw score scale with just a few exceptions.

➤ *Equating differences in scaled scores between repeater and total groups*

Figures 11-15 display the scaled score differences between equatings in the repeater group and in the total group by new-form raw score levels. Similar to the above findings for Figures 6-10, we found many of the scaled score differences practically significant for each of the study equatings but most of the practically significant differences were not statistically significant, especially for A-V, A-Q, B-V, and B-Q. In addition, the directions of equating differences shown in Figures 11-15 also support the notion that the repeaters were at least as able as (and often more able than) the first-time examinees on the study exam.

➤ *Equating differences in scaled scores between first-timer and total groups*

Figures 16-20 compare the scaled score outcomes based on equatings in the first-timer group and in the total group. Comparisons between total group equating and first-timer-group equating is commonly used to demonstrate effects of repeater performance



on score equating, as the total group includes repeaters and the first-timer group excludes repeaters from equating. Overall, Figures 16-20 show no significant differences in scaled scores between the total group and first-timer group (neither statistically nor practically) across various study equatings/administrations, which implies insignificant effects of repeater performance on score equating for the study exam.

➤ *RMSD results*

The *RMSD* results presented in Figures 21 to 25 indicate that equating differences in scaled scores between the subgroups and total group were generally not significant (neither statistically nor practically) in the middle region of the new-form raw score scale, but at the two ends of the raw score scale the differences tended to be significant (both statistically and practically), indicating substantial invariance problems for the total-group scaled scores for the low- and high-achieving examinees.

➤ *Practical vs. statistical significance of scaled score differences*

In general, many of the scaled score differences shown in Figures 6-25 were practically significant while not as many were statistically significant. This is especially true for Figures 6-15 involving the repeaters, which had rather large standard error bands around the scaled score differences. For Figures 16-20 involving the first-time examinees, though, the scaled score differences were largely not of practical or statistical significance. And, the *RMSD* results in Figures 21 to 25 indicated that scaled score differences in middle region of the new-form raw score scale were neither statistically nor practically significant, whereas the differences at one end or both ends of the scale tended to be practically significant but not statistically significant.

Overall, these results seem to suggest an effect of significance evaluation criterion/measure. While we usually do not expect results of practical and statistical significance to agree with each other, findings in this study seemed to imply that the practical evaluation criterion might be too sensitive in detecting significant scaled score differences whereas the statistical criterion might not reveal scaled score differences of importance to score fairness especially when available sample sizes were small as with the repeater case in this study. We will further discuss issues for the practical evaluation criterion in the Discussion section.



➤ *Where did the scaled scores differ?*

Different than the *RESDs* and *REMSDs* in Table 8, the positive and negative scaled score differences exhibited in Figures 6-25 further convey that the scaled scores based on the repeater-group equating were not consistently higher or lower than those for the total-group equating or first-timer-group equating across various forms and measures. Usually significant scaled score differences (either practically or statistically) occurred at the lowest and highest regions of the new-form raw score scale, especially for scaled score differences between the repeater group and the other two groups presented in Figures 6-15.

Across various forms and types of comparison, some of the scaled score differences at the two ends of the raw score scale were perfect zero; most of which were the consequence of truncating the equated scale scores that were out of the scale range of reported scores. Specifically, the out-of-range scaled scores were rounded to the possible min/max (i.e., 20/80) of the reporting scale. The reasons for such truncation/rounding were explained previously in the Method section (see the subsection for “*A Focus on Raw-to-Scale Equating*” and Footnote #3). As out-of-range scaled scores were truncated to the same scaled score values (scaled scores lower than 20 were rounded to be 20, and scaled scores higher than 80 were rounded to be 80) for equatings in various groups/subgroups, equating differences in the very low and very high scaled score regions were likely to become zero, indicating subpopulation invariance.

Consistency across Summary Statistics in Invariance Outcomes

Some aspects of Figures 6-25 reflect the same issues of invariance measures and data suggested in the summary statistics of Table 8. In particular, the first-time examinees comprised more of the total group than the repeaters, so that the equating differences in scaled scores between the first-timer group and total group were relatively small (Figures 16-20) compared to the scaled score differences between the repeater group and total group (Figures 11-15). The standard errors involving the scaled scores based on the first-timer-group were also smaller than those for the repeater-group.

For the Verbal and Quantitative forms with statistically significant repeater *RESDs* in Table 8 (A-Q, B-V, and B-Q), the scaled score differences exhibited in Figures



12-14 were practically significant at certain new-form raw score levels but they were hardly statistically significant because of the large standard error bands associated with the small repeater sample sizes. For the Quantitative forms with statistically significant first-timers RES_{D_jS} in Table 8 (A-Q and B-Q), Figures 17 and 19 showed that the small scaled score differences were not statistically or practically significant at various new-form raw score levels, except for those at the relatively low end of the raw score scale (specifically, at the raw score levels of 11 and below for A-Q, and 13 and below for B-Q). In addition, Figures 11, 12 and 15 show that some of the scaled score differences between the repeater-group equating and total-group equating at new-form raw score levels were practically significant (for forms A-V, A-Q and C-Q), whereas the corresponding RES_{D_jS} and $REMSD$ statistics in Table 8 suggested otherwise.

While most of the standard error criteria in Table 8 appeared more stringent (i.e., smaller bands) than the DTM criteria of 0.5 point, in Figures 6-15 the standard error bands are considerably wider than the DTM bands. The scaled score differences in Figures 6-15 show how the equatings in the repeater group differed than the equatings in the other study groups (i.e., the first-time examinee group and total group). The wider standard error bands reflect larger variability of the summary statistics (i.e., the simple scaled score differences) exhibited in Figures 6-15, which was primarily due to the small sample sizes at the raw score levels. However, the standard error bands are quite close to the DTM bands in Figures 16-20, and in Figures 21-25 the standard error bands are usually narrower than the DTM bands at the middle raw-score levels but wider than the DTM bands at the two ends of the raw-score scale, which is a reflection of the new-form sample distributions (more examinees in the middle and fewer at the two ends).

Major study findings are summarized below--

- The self-reported repeater data used in this study was verified empirically; the data looked reasonably sound and was the best option available for this study.
- General trends of repeater performance include: Fairly stable repeater scores across testing occasions; High-performing examinees might not improve their scaled scores by retesting as the low-performing examinees could; on average, repeaters were at



least as able as the first-time examinees on the study exam (they were actually more able than the first-time examinees on half of the study tests). However, further studies are needed to better understand repeater performance patterns while taking into account repeater demographic background information.

- Overall, the total-group equating function and its resulting scaled scores seemed reasonably invariant across subpopulations. In general, differences in scaled scores between the total-group and first-timer-group equatings were negligible from both practical and statistical perspectives. This implies that effects of repeater performance on score equating was not significant for the study exam.
- Despite a few mixed results based on different invariance measures, overall the repeater-group equating seemed different from the equatings in the first-timer group and total group. Although many of the differences in scaled scores were practically significant, most of the practically significant differences were not statistically significant.
- Large equating differences on the scale of reported scores often occurred at the two ends of the new-form raw-score scale. Although the differences were usually of practical significance, they were seldom significant statistically.
- There might be a criterion/measure effect for the significance evaluation of equating differences on the scale of reported scores—the practical evaluation criterion might be too sensitive in detecting significant scaled score differences, whereas the statistical evaluation might be limited by small study sample sizes and as a consequence not reveal scaled score differences of importance.
- The invariance outcomes based on various summary statistics looked consistent overall, despite some discrepancies associated with the small repeater groups. Differences in invariance results suggested by different measures could be reasonably explained.

Discussions

We discuss important study findings and their implications on test equating, as well as study limitations in this section.



Effects of Repeater Performance on Score Equating

Overall, for both Verbal and Quantitative measures we found the differences between total-group and first-timer-group equatings negligible from either a practical or a statistical perspective, which implies negligible repeater effects on score equating for the two measures of the study exam. In addition, equating in the repeater group generally looked different from the equatings in the first-timer group and total group. However, despite their practical significance the equating differences were not statistically significant most of the time (partly due to the small repeater sample sizes available for this study). Therefore, it seems safe to conclude that for the study exam there was not a significant repeater effect on score equating.

Given the conclusion of non-significant repeater effect for the study exam, it seems that the testing program could consider using the total-group equating (instead of the first-timer group equating) to enhance the precision of equating by increasing the equating sample sizes. Nevertheless, to be prudent the practically but non-statistically significant equating differences between the repeater-group equating and the equatings in the first-timer group and total group deserve to be further scrutinized. Although these differences were not statistically meaningful, their practical meaningfulness and the consistent observations of such differences across study administrations and measures seem to support the uniqueness of the repeater-group equating. Perhaps more studies with larger repeater sample sizes could be conducted to see whether the differences remain statistically non-significant. Efficacy of the practical criterion for evaluating the significance of equating differences should also be re-assessed (discussed below).

Practical Criterion for Evaluating Equating Differences

It is common to have statistical and practical significance evaluation outcomes that do not agree with each other. Nevertheless, our study findings seem to suggest that the practical evaluation criteria (i.e., the DTMs) might be too sensitive in detecting important equating differences. While there were a large number of equating differences at new-form raw score levels that were practically significant, many of these differences were not of statistical significance. Although the small subgroup sample sizes for this study could account for part of the statistical significance evaluation outcomes, the



pronounced disparities between the practical and statistical evaluation outcomes did cast doubt on adequacy of the practical evaluation criteria. As discussed previously in the Method section, the selected DTMs on the reported score scale looked reasonable in the context of cut-score decisions with the rounding practice, but their efficacy could be limited by their own arbitrary and subjective nature.

Unrounded vs. Rounded Scores for Evaluating Practical Significance

Because rounded integer scaled scores were reported to candidates taking the study exam, which could be directly compared to some admissions screening threshold, it seemed reasonable to use rounded scaled scores to evaluate the practical significance of equating differences in order to be more consistent with the score reporting practice. However, in this study we focused the comparisons of scaled score differences resulting from different equatings on unrounded values to avoid potential confounding effects due to rounding, which could dramatically change the evaluation outcomes. With rounded scores, some of the significant equating differences could be due to rounding instead of equating.

Choice between unrounded and rounded scores primarily depends on whether the advantages outweigh the disadvantages. In an effort to assess the efficacy of rounded and unrounded scores, we compared the outcomes of equating differences based on the rounded scores⁵ to those based on the unrounded scores. Overall patterns of equating differences based on the rounded and unrounded scores could look rather consistent, but rounding could result in equating differences that were much larger or smaller than they actually were, depending on the values of corresponding unrounded scores. Sometimes rounding could make unrounded differences from -0.49 to 0.49 zero (e.g., both 40.49 and 40.00 could be rounded to 40), but other times it could make very small differences large if the unrounded scores were on or near the 0.5 rounding boundary (e.g., 40.50 could be rounded to 41 while 40.49 was rounded to 40). Based on this and the fact that in an invariance study we care more about scaled score differences due to equating than rounding, it seems more appropriate and important to compare equating differences using

⁵ Numerical values of the practical significance criteria in this study changed from +/- 0.5 to +/- 1 (which widened the band for practical invariance) when the rounded scaled scores were used instead of the unrounded scaled scores.



unrounded scores than using rounded scores—even when the goal is to evaluate "practical" significance of equating differences. We could never be sure how rounding affects the results of equating differences based on rounded scores, unless we look into the results based on unrounded scores.

Validity of the Self-reported Repeater Information

In testing practice, it is often not feasible or easy to identify repeaters because there is usually not a built-in mechanism in the scoring system to automatically identify repeaters and merge their score records across administrations, and the post-administration analysis window is often too narrow to allow sufficient time for testing programs to manually match empirical data across administrations to identify repeaters. As a result, repeaters are usually identified on a self-reported, voluntary basis. A common way to solicit self-identified repeater information is to employ an examinee survey at the registration or testing time. Despite the ease and convenience of the survey approach, information collected through the voluntary response of examinees is not likely to fully reflect the actual repeater status of examinees. Reliability and validity of the survey results are usually a cause for concern, which is especially true when examinees have a motivation to conceal their repeater identity (e.g., examinees may want to distance themselves from their poor test scores from before). Furthermore, repeater surveys are often not designed to acquire sufficient repeater information in consideration of examinees' ability to recall their test-taking history and their willingness to respond to a lengthy survey.

In this study we used the self-reported but empirically verified repeater information to facilitate our analysis of repeater effects, which was the best option we could have based on the available data. The verification process was labor-intensive and time-consuming but it helped to raise our confidence in the self-reported repeater data. If we used the empirically matched examinee records across administrations for our analyses instead, we would under-identify repeaters on the study exam because the empirical repeater data was restricted by a lack of effective matching variables for aggregating cross-administration data for the same examinees.

Nevertheless, if we could have a better way to identify repeaters or to more thoroughly verify the self-reported data we would have more confidence in the resulting



study findings. Therefore, we recommend the use of more reliable repeater information (either through effective identification or verification) for future research to ensure the quality of study data.

Invariance of Equating on Verbal vs. Quantitative

The effects of repeater performance on score equating for Verbal and Quantitative were quite similar based on the findings of this study. Because operationally the Verbal subscore is weighted more (by 10%) than Quantitative subscore in calculating the composite score, subpopulation invariance property of the Verbal equating would have a stronger impact on scaled composite score outcomes than that of the Quantitative equating. That is, if equating is ever not invariant across subpopulations its impact on the reported composite score would be more serious when the equating is for the Verbal measure than when it is for Quantitative.

Overall vs. Specific Repeater Effects

Primarily due to limited study sample size, we only examined the overall, non-specific repeater effects without differentiating repeaters by the number of retakes, the time interval between testing occasions, etc. We also did not consider atypical cases when examinees repeated the same test or anchor test, because based on our experience the frequency of such cases was very low. In addition, in normal testing situations we would strive hard to prevent or minimize any security or fairness problems associated with the reuse of items or test forms, even before equating takes place.

Investigation of the overall, non-specific repeater effects is regarded as the first step toward unveiling the potential impact of repeater performance on equating and scoring. And, in-depth examination of general repeater performance patterns would shed some light for future research. To better untangle the source of repeater effects on score equating, more research that focuses on specifically defined repeater subgroups is needed.

Range Restriction due to Self-Selected Repeaters

Repeater population is usually restricted in range because of the self-selection nature of repeaters. Consequently, performance of repeaters can have a significant impact on equating outcomes when repeaters are included in equating samples. There is a common belief that examinees who did not meet the required selection criterion or



passing standard in prior test administration are more likely to repeat the exam; in such case, repeaters tend to be less able than the first-time examinees. Nevertheless, for some exams used for selection/admissions purposes repeaters are not necessarily less able than the first-time examinees (such as the repeater group in this study), because the general examinees (not limited to the low-achieving ones) have an incentive to achieve a higher score to enhance their chance for advancement or admissions to a more prestigious institution. In short, the nature and extent of range restriction resulting from the self-selection of repeaters can vary from testing program to program. Therefore, while considering effects of range restriction on score equating we first need to have a good understanding for range restriction that is pertinent to a particular testing program.

Limitation due to Raw-to-Scale Equating

Aside from the already mentioned limitations of this study (e.g., imperfect self-reported repeater data and insufficient subgroup sample size for further analyses), we also faced a trade-off between equating practicality and precision. In this study, we chose to focus on the raw-to-scale equating for its practical importance to score fairness. However, the raw-to-scale equating function depends on the previously established reference-to-scale equating function, which might be subject to effects of repeater performance unless the effects were controlled for equatings in the past. As a result, as much as the repeater effects were controlled for the raw-to-raw equating in this study, the study outcomes based on the new-to-scale equating could not be free of potential bias due to the prior reference-to-scale equating. Nevertheless, by focusing on the scaled score outcomes involving additional computations, transformations and truncations on top of the raw-to-raw equating results, this study addressed the repeater effects on scaled score conversions which are seldom investigated because of their complexity. In essence, differences in scaled scores in this study reflect not only the differences in raw-to-raw equatings but also the characteristics of reference-to-scale conversions, as well as the treatment to equating outcomes for score reporting purposes (e.g., rounding and truncation).

It is important to study effects of repeater performance on score equating and the study outcomes should guide the design and shape the strategies of future equating. As characteristics of repeater performance and equating outcomes are intricately related, we



need to compliment the equating study with information on general/specific performance trends of repeaters. And, more research dealing with data of varying repeater characteristics from a variety of testing programs are needed to provide a broad range of evidence of repeater effects on equating, which should enhance equating practice and inform critical program decisions on scoring fairness.



References

- Andrulis, R. S., Starr, L. M., & Furst, L. M. (1978). The effects of repeaters on test equating. *Educational and Psychological Measurement, 38*, 341-349.
- Cope, R. T. (1986). Use versus nonuse of repeater examinees in common item linear equating with nonequivalent populations (ACT Technical Bulletin 51). Iowa city, IA: American College Testing.
- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). The chain and post-stratification methods for observed-score equating: Their relationship to population invariance. *Journal of Educational Measurement, 41*, 15-32.
- von Davier, A. A., & Wilson, C. (2008). Investigating the population sensitivity assumption of Item Response Theory true-score equating across two subgroups of examinees and two test formats. *Applied Psychological Measurement, 32*(1), 11-26.
- Dorans, N.J. (2004). Using the subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41*(1), 43-68.
- Dorans, N. J., & Holland, P. W. (2000). Population invariance and equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*, 281-306.
- Dorans, N. J., Liu, J., & Hammond, S. (2008). Anchor test type and population invariance: An exploration across subpopulations and test administrations. *Applied Psychological Measurement, 32*(1), 81-97.
- Gorham, J. L. & Bontempo, B. D. (1996). *Repeater patterns on NCLEX® using CAT versus NCLEX® using paper-and-pencil testing*. Paper presented at the annual meeting of the American Educational Research Association, New York.
- Harris, D. J. (1993). *Practical issues in equating*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta.
- Holland, P. W. (2003). Overview of population invariance of test equating and linking. In N. J. Dorans (Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement Program examinations* (ETS RR-03-27, pp. 1-18). Princeton, NJ: Educational Testing Service.



- Kingston, N., & Turner, N. (1984). *Analysis of score change patterns of examinees repeating the Graduate Record Examinations General Test* (ETS RR-84-22). Princeton, NJ: Educational Testing Service.
- Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag.
- Liu, J., Cahn, M. F., & Dorans, N.J. (2006). An application of score equity assessment: invariance of linkage of new SAT to old SAT across gender groups. *Journal of Educational Measurement*, 43(2), 113-129.
- Liu, M., & Holland, P. W. (2008). Exploring population sensitivity of linking functions across three Law School Admission Test administrations. *Applied Psychological Measurement*, 32(1), 27-44.
- Moses, T. P. (2006). *Using the kernel method of test equating for estimating the standard errors of population invariance measures* (ETS RR-06-20). Princeton, NJ: Educational Testing Service.
- Yang, W. (2004). Sensitivity of linkings between AP multiple-choice scores and composite scores to geographical region: An illustration of checking for population invariance. *Journal of Educational Measurement*, 41(1), 33-41.
- Yang, W., & Gao, R. (2008). Invariance of score linkings across gender groups for forms of a testlet-based College-Level Examination Program examination. *Applied Psychological Measurement*, 32(1), 45-61.
- Yi, Q., Harris, D. J., & Gao, X. (2008). Invariance of equating functions across different subgroups of examinees taking a science achievement test. *Applied Psychological Measurement*, 32(1), 62-80.
- Zhang, Y. (2008). Repeater analysis for TOEFL iBT (ETS RM-08-05). Princeton, NJ: Educational Testing Service.