Listening. Learning. Leading.®

Research Report
ETS RR–11-17

Repeater Effects on Score Equating for a Graduate Admissions Exam

Wen-Ling Yang

Andrea M. Bontya

Tim P. Moses

April 2011

Repeater Effects on Score Equating for a Graduate Admissions Exam

Wen-Ling Yang, Andrea M. Bontya, and Tim P. Moses
ETS, Princeton, New Jersey

April 2011

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Technical Review Editor: Dan Eignor

Technical Reviewers: Mary Grant and Mei Liu

Abstract

Using self-reported but empirically verified repeater groups, we analyzed vast amounts 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 nonequivalent groups with anchor test (NEAT) design. Both linear and nonlinear equating models were considered in deriving the equating functions for various study groups. We evaluated scaled score differences between equating 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 the 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 exam being studied. 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 is 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 of the practical criterion for evaluating the significance of equating differences, and study limitations are also discussed.

Key words: score equating, significance of equating differences, repeater effects, equatability

Acknowledgments

This report is based on research results presented at the annual meeting of the American Educational Research Association, San Diego, CA, April 2009. The authors would like to thank Marna Golub-Smith, Daniel Eignor, Michael Kane, Anna Kubiak, Amy Schmidt, Mary Grant, and Mei Liu for their insightful comments and suggestions on earlier versions of this paper.

Table of Contents

List of Tables

Page

List of Figures

Page

Score equating is commonly used for ensuring comparable scores across different test 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 scaled 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 arises over the practice of excluding repeaters from the 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 (Gorham & Bontempo, 1996; Kingston & Turner, 1984; Zhang, 2008). And, changes in scaled scores, ability estimates, and/or passing rates were often the unit of analysis, despite the fact that equating was critical in deriving the scaled 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 groups design with anchor items (assuming two equally reliable tests) and evaluated the repeater effects in terms of differences in the resulting equating parameters, cutoff 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 nonequivalent groups with anchor test (NEAT) design. Cope compared the results of linear equating models using examinee data with and without repeaters and used equating chains to evaluate the relative accuracy of various equating outcomes. Because the equating outcomes based on the first-time examinees were not necessarily or substantially more accurate than the equating outcomes based on the total examinee group, and the relative accuracy of equating seemed to depend partly on the specific linear equating method used, Cope had reservations about the practice of routinely excluding repeaters from equating for the test being studied. As a result, the author 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 effects 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 equating 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 studies have 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 equating 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 stakes.
- 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, nonspecific 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—NEAT design.
- Equating methods—Both linear and nonlinear models.
- Unit of analysis—Equating outcomes expressed on the reporting scale (i.e., scaled 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 to make recommendations about the treatment and 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 more specific to the testing program being studied, but would also benefit other testing programs alike:

3. To delineate the general patterns of repeater rates and trends for the exam being studied and to evaluate the soundness of the program's operational practice of excluding repeaters from the equating samples.
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 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 a paper-and-pencil test format, the exam being studied 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 scaled scores valid for 5 years, examinees (even those who scored high previously) have an incentive to retake the exam 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 practice effects, while scores on different test forms are equated operationally using the 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 7 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 5 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 to get 5 years' worth of score data was based on the assumption that an examinee who took the exam more than 5 years ago was less likely to repeat the exam being studied and, if the examinee did repeat the exam, he or 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 SEEs and the 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 for the Verbal measure.[1] In general, there were 65 items in a Quantitative test form and 90 items in a Verbal form. However, actual test length varied due to the removal of items with poor performance before equating/scoring. For each of the "new" forms analyzed in this study, Table 1 shows the possible score ranges for the total test and 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 shows the score means and standard deviations (as percentages of possible maximum score points) on both total and anchor tests, as well as the correlation coefficient between the total and anchor test scores. Table 2 presents similar information for the study forms analyzed as reference forms. The *V* or *Q* in the form name indicates whether a test form is for Verbal or Quantitative, and the *R* in the form name refers to the reference form.

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 10 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 group comparison outcomes in detail later in the Results section.

Also shown in Tables 1 and 2 were the anchor-total correlation coefficients across test forms and examinee groups, which ranged from 0.73 to 0.93. While the correlation coefficients across study forms looked quite different, differences in correlations across examinee groups were very small, except for Form RB-Q (for which the correlation for the repeater group was much lower than those for the other two groups). Such differences in anchor-total correlation might lead to different levels of equating efficacy across forms (or across examinee groups for Form RB-Q).

A close inspection of the raw mean percentage scores (see Tables 1 and 2) also indicated that Verbal means were more consistent across forms than Quantitative means on both 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 Verbal forms constructed were more comparable to each other than 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 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 who 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 nonspecific 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 first-time repeaters, the second-time repeaters, the third-time repeaters, and so on) would not support meaningful statistical analyses. 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 6 months, 1 year, 2 years) and the ability levels of repeaters. Therefore, in this study we focused on examining the effects of the overall, nonspecific 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 distance themselves from the previous records 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, and birth dates.

General Trends in Repeater Performance

It is important to grasp the general trends in repeater performance prior to considering repeater effects on score equating. To do so, we computed test-retest correlation coefficients using scaled (equated) score data in the repeater group and investigated repeaters' scaled score gain/loss. Using raw score data, we also compared the performance of repeaters to that of first-time examinees on each of the study forms.

Test-retest correlation. Since some examinees had repeated the exam more than once, to standardize the selection of test scores and to take into account data recency we focused on the two most recent scores of individual repeaters while looking into the test-retest relationship. Because scores of examinees in the overall, nonspecific repeater group were from a broad range of exam administrations, we used the scaled scores that were comparable across testing occasions for calculating the test-retest correlation coefficients.

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

Scaled score gain/loss. In addition to examining general scaled score gain/loss for the overall, nonspecific repeater group, we also compared patterns of scaled score gain/loss across various repeater subgroups that differed in their prior performance. Conditional distribution data for repeaters (i.e., percentages of repeaters conditioned on their prior test performance) on study forms were used for this analysis. By accounting for repeaters' prior test performance, we could gain an in-depth understanding of the repeater scoring trends.

Furthermore, to mitigate potential regressive effects resulting from aggregating repeaters across administrations in forming the overall, nonspecific repeater group, we analyzed repeater score gain/loss patterns with a more narrowly defined but much smaller repeater sample, namely, the repeaters who took study forms A-V (for Verbal) and A-Q (for Quantitative).

Comparing repeater performance to first-timer performance. To study repeater performance that was not influenced by equating practice and its effects on scoring consequences, for each of the target equating forms, we also compared the performance of the repeaters to the performance of the first-time examinees using their *raw* total scores on the same test form.[2] 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-sample Z test as follows:

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

$$\textit{where} \quad \sigma^2_{\bar{X}_r} = \frac{\sigma^2{}_r}{n_r} \quad \textit{and} \quad \sigma^2_{\bar{X}_{nr}} = \frac{\sigma^2{}_{nr}}{n_{nr}}.$$

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 these 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 nonlinear equating models. Specifically, for each of the equating relationship being studied, 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 the data of a particular group. 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 equating outcomes 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 a subgroup 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 the raw-to-scale equating that converts new-form raw scores to scaled scores used for score reporting, instead of the raw-to-raw equating that converts new-form raw scores to reference-form raw scores. Consequences of the raw-to-scale equating are much more critical to score fairness in practical equating situations.

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) equating function and the reference-to-scale scaling function, which not only adds complexity to the equating process but can also complicate the evaluation of equating outcomes. For instance, special consideration/treatment is needed for determining the scaled score values for equated raw scores that go beyond the reference-form possible score range (i.e., when impossible scaled score conversions occur).[3] The method used for combining the equating and scaling functions may also affect final scaled score outcomes. The reliance on the reference-to-scale scaling 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 the details of equating differences across various study groups (in scaled score units) by new-form raw score levels using score plots. These graphical presentations help to show the direction and magnitude of equating differences along the new-form score scale.

Also, using the set of equatability indices (also known as score equity assessment or SEA indices) developed for checking the subpopulation invariance properties of an equating function and for checking the equity of scores across subpopulations (Dorans, 2004; Dorans & Holland, 2000; von Davier, Holland, & Thayer, 2004; Yang, 2004), we summarized the comparison outcomes between the total-group equating function and the equating for the total group's respective subgroups (i.e., the repeater and the first-time examinee groups). Used widely in a series of studies since 2000 (Dorans, Liu, & Hammond, 2008; Liu, Cahn, & Dorans, 2006; Liu &

Holland, 2008; von Davier & Wilson, 2008; Yang & Gao, 2008; Yi, Harris, & Gao, 2008), the set of equatability indices helps to assess the overall adequacy of the total-group equating function or the first-time-examinee group equating function.

Specifically, the summary statistics we used include the root mean square difference (*RMSD*), the root expected square difference ($RESD_j$), 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 $RESD_j$ 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 $\Sigma w_j=1$ (Dorans, 2004; Dorans & Holland, 2000; von Davier, Holland, & Thayer, 2004). As with *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:

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

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 $\Sigma w_{xP}=1$. 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 the new-form administration.

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

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

In addition to the statistics described above, we used statistics of simple differences to summarize scaled score differences between equating 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 exam being studied. Specifically, the criterion for evaluating practical significance of scaled score differences was based on half a score point on the subscore scale of the exam.

In addition to reporting the composite score, which is the sum of the weighted subscores for the four component measures, the exam also reports a 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 place 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 exam, it at first seems appropriate to say that in general a 1-point difference on the 20-to-80 subscore scale is a DTM to examinees taking the exam and test users, because graduate institutions often apply a cutscore to screen their applicant pools, and 1-point difference on the subscore scale could translate to several points on the composite score scale for the study exam. However, because operationally half a score point on the subscore scale would be rounded to 1 for score reporting purpose, it actually seems more appropriate to define the DTM as half a score point on the subscore scale. From a practical perspective, we would consider an equating difference negligible if it is smaller than this DTM.

Simulations for Standard Error Estimation

To estimate SEEs, 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 forms[4]) that were used for the smoothed chained equipercentile equating 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, of the scaled score differences, and of the 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 for the repeater groups, it was crucial to evaluate the statistical significance of equating differences (on the scale of reported scores) to determine whether study findings were more than 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

In this section, we first present the identification and verification outcomes for the data containing self-reported repeaters to set the grounding for this study. Then, we present analysis outcomes showing the general trends in repeater performance to put the study equating in proper perspective and to aid in the interpretation of study findings that follow. Following a description of the equating for various study forms, we present the results of repeater effects on score equating based on various statistics. Lastly, we compare results of various invariance measures. A highlight of major study findings is provided at the end of this section.

Identification and Verification of Repeater Status

Overall, we found nearly an 88% match (72% nonrepeaters 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. However, the self-reporting approach identified the other 11% or so examinees as repeaters while the empirical approach did not agree, which was likely to be a miss by the empirical approach because of the imperfect matching of examinees' records across administrations. Although the empirical approach picked up some examinees as repeaters while the self-reporting approach failed to do so, the percentage was rather small—only about 0.5%. 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 pick up those not identified by the self-reporting approach (i.e., those examinees who concealed their repeater identity in the voluntary repeater survey).

The disagreement in repeater identification outcomes between the empirical approach and 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 real repeaters, we would avoid using the voluntary, self-reported repeater information for our study. However, none of the available matching variables, or any 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 for this study. Therefore, we used the self-reported repeater data for our analyses to avoid under-identification of real repeaters. In general, the self-reported repeater data looked reasonably sound. The data might not be perfect, but it appeared to be the best option we could have for this study.

General Trends in Repeater Performance

Across various study administrations, most of the examinees in the general, nonspecific (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, empirical findings on the number of retakes still offered useful insights for studying general repeater trends and patterns, especially when the self-reported repeater data did not provide such information at all (the repeater survey of the exam being studied was not designed to collect such information).

Test-retest correlations. For the overall, nonspecific repeater group (N = 6,256), the test-retest correlation coefficient was 0.74 for Verbal and 0.72 for Quantitative, based on individual repeaters' two most recent scaled scores. 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 care is 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, and results may not be generalized to the entire examinee population.

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

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	-
Verbal overall ($0 < t < 7$)		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	-
Quantitative overall ($0 < t < 7$)		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.

Overall, for both of the 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 (years). The 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. These findings are not in agreement with the common belief that r would decrease when *t* increases (i.e., the practice and/or recency effect is likely to diminish as time goes by). Perhaps in the 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.

Repeater score gain/loss. An inspection of the scaled scores of the overall, nonspecific repeater group revealed that:

- For Verbal, close to 59% of the repeaters improved their scores after retaking the exam, more than 35% had score decreases, and about 6% saw no score change.
- Findings for Quantitative were very similar to those for Verbal with slight differences.
- On average, the repeaters improved their scaled scores by only 2.2 points or so on either Verbal or Quantitative, but the variability of scaled score changes was very large—the standard deviation for either one of the measures was about 7.2, suggesting a cancellation of large score gains and losses due to summing and averaging, which was not uncommon for score data spanning a large number of administrations.

The distributions of repeater score gain/loss conditioned on prior test performance provided an in-depth look at the trends in repeater performance. They allow comparisons of repeater score gain/loss across various repeater subgroups that differed in prior test performance. Table 4 shows the conditional distributions for the overall, nonspecific repeater group. And, Table 5 presents the conditional distributions for the more narrowly defined but much smaller repeater group on new forms A-V and A-Q.

Table 4

Conditional Scale Score Gain/Loss for the Overall, Nonspecific Repeater Group

Test measure	Scale score on the prior test	*N*	% of examinees with score gain/loss (later test score - prior test score)									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	7	14	14	14	43	0	14	0	0	0	65	-8
	61-70	229	4	5	21	37	4	22	7	0	0	61	-3
	51-60	1,559	2	4	14	29	7	27	13	3	1	54	0
	41-50	2,452	1	3	9	21	6	29	20	10	3	48	2
	31-40	1,474	0	2	7	18	4	29	22	13	5	40	4
	20-30	535	0	0	4	12	6	25	25	19	10	33	7
Quantitative	71-80	90	8	4	19	23	10	26	10	0	0	71	-3
	61-70	607	2	7	14	28	6	24	14	5	0	63	-1
	51-60	1,834	1	4	13	24	6	28	16	6	2	56	1
	41-50	2,522	0	2	10	21	6	26	21	9	4	48	3
	31-40	1,134	0	0	3	16	5	30	25	14	7	42	5
	20-30	69	0	0	0	6	3	20	30	26	14	38	9

Note. Scale scores from two most recent testing occasions of 6,256 repeaters were used for this analysis. 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, because examinees in the overall, nonspecific repeater group were from a wide range of exam administrations.

Table 5

Conditional Scale Score Gain/Loss for Repeaters on New Forms A-V and A-Q

Test form (measure)	Scale score on the prior test	*N*	% of examinees with score gain/loss (Form A score - prior test score)									Average scale score on Form A	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		
A-V (Verbal)	71-80	-	-	-	-	-	-	-	-	-	-	-	-
	61-70	13	8	8	0	31	8	31	15	0	0	62	-1
	51-60	86	2	1	20	28	7	24	7	8	2	54	0
	41-50	178	1	2	6	20	10	25	25	10	1	48	3
	31-40	101	0	1	4	14	5	32	24	18	3	42	5
	20-30	27	0	0	7	15	0	7	11	44	15	35	9
A-Q (Quantitative)	71-80	6	0	17	33	33	17	0	0	0	0	70	-6
	61-70	31	3	16	16	29	0	29	6	0	0	61	-3
	51-60	123	2	4	20	28	5	24	12	3	0	54	-1
	41-50	170	1	3	15	28	5	25	18	2	3	46	0
	31-40	70	0	0	3	19	9	33	24	11	1	42	4
	20-30	5	0	0	0	0	0	40	60	0	0	35	6

Note. Out of the 537 self-identified repeaters on Form A, we were able to retrieve scale scores from the preceding testing occasion for 405 repeaters by matching score records across exam administrations. Scale scores on Form A and the preceding tests of the 405 repeaters were used for this analysis. The preceding tests might not be the same for different repeaters.

The last column in Table 4 shows the average scaled 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 more informative than the average scaled score change of 2.2 for the overall repeater group. For both Verbal and Quantitative, data in the last column of Table 4 also indicate that in general the higher the repeaters' scores on the prior test, the lower their score gains on the later test. There were even negative score gains (i.e., score losses) for the repeaters scoring above 60 on the reporting scale for both measures. This result looked reasonable and was not surprising because of both the ceiling and regression to the mean effects. The average scaled scores on the later test, as reported in the second last column of Table 4, looked consistent with this observation. It implies that the high-performing examinees may not increase their scaled scores by retesting (the retest scores could be even lower than before), whereas the low-performing examinees could improve their scaled scores by a substantial number of points.

At the center of Table 4 are columns showing the percentages of examinees with different degrees of score gain/loss conditioned on the examinees' prior test performance. Results of the conditional score gains/losses, especially those shown in the three shaded columns in the middle, further indicated that regardless of prior test performance (except for the worst performers on the prior test) in general the majority of repeaters had a score change of no more than 5 points. However, the lower the prior test scores, the more repeaters with score gains of more than 5 points (and the fewer repeaters with score losses of more than 5 points). This was consistent with the previous findings based on the last two columns of Table 4.

The conditional distribution outcomes based on the repeaters taking new forms A-V and A-Q (see Table 5) were largely consistent with the results for the overall, cross-administration repeater group, despite its much smaller sample size.

Performance of repeaters versus first-time examinees on new forms. Analyses of repeater performance based on reported scaled 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 scaled scores across administrations. Therefore, we also analyzed repeater performance by comparing the raw scores of various examinee groups on the same test form, which were not influenced by subsequent equating yet.

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); 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).

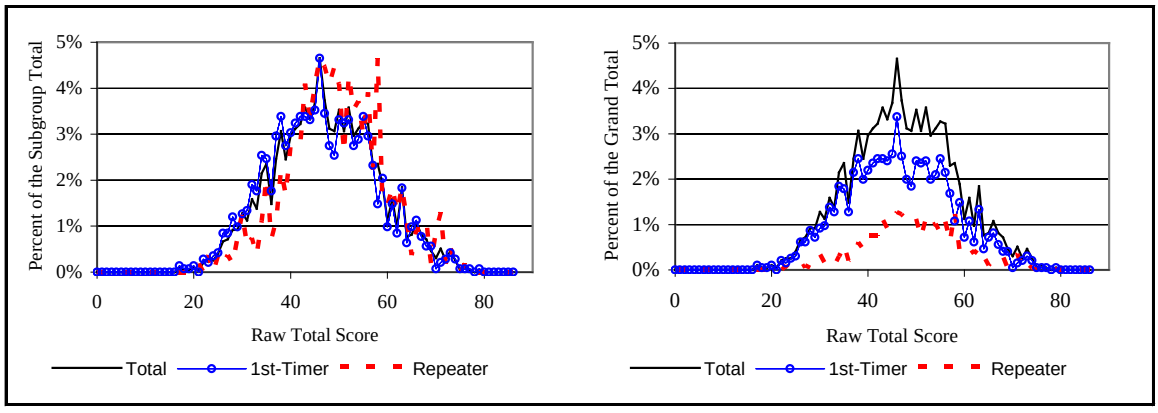

Figure 1. **Observed relative frequency distributions for Form A-V.**

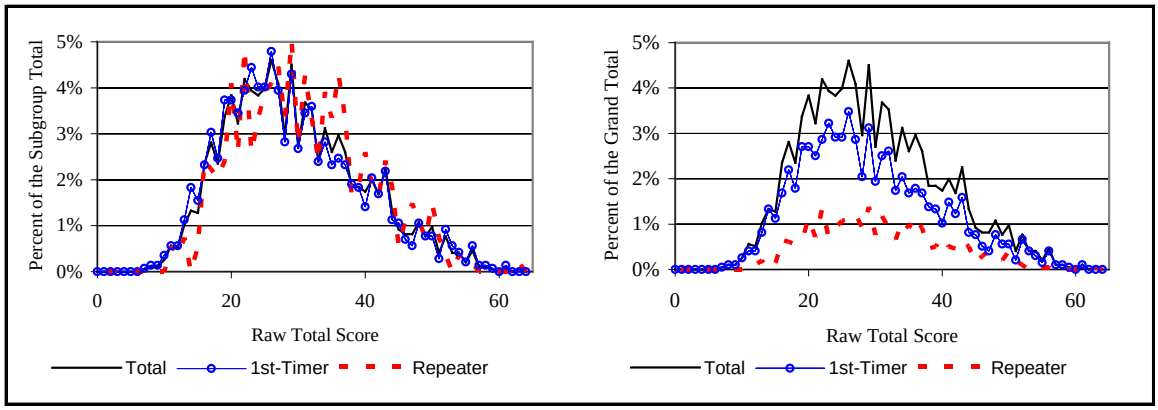

Figure 2. **Observed relative frequency distributions for Form A-Q.**

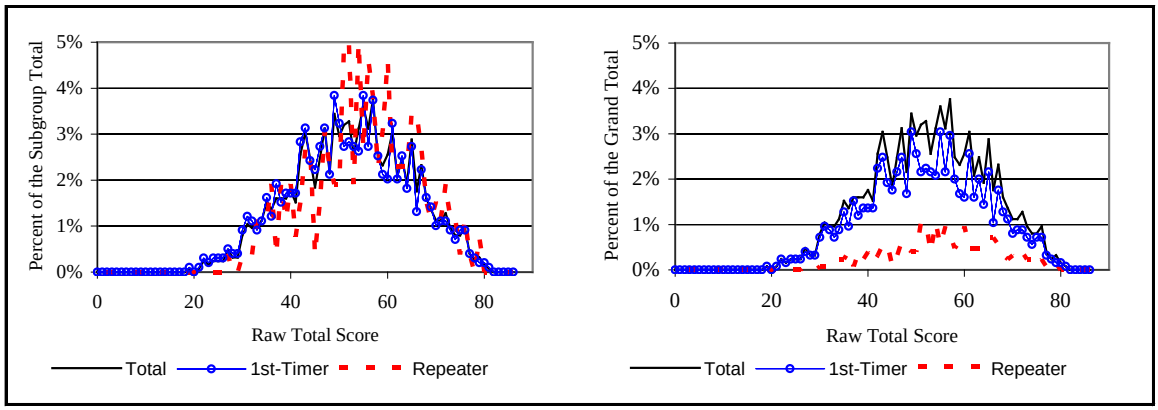

Figure 3. **Observed relative frequency distributions for Form B-V.**

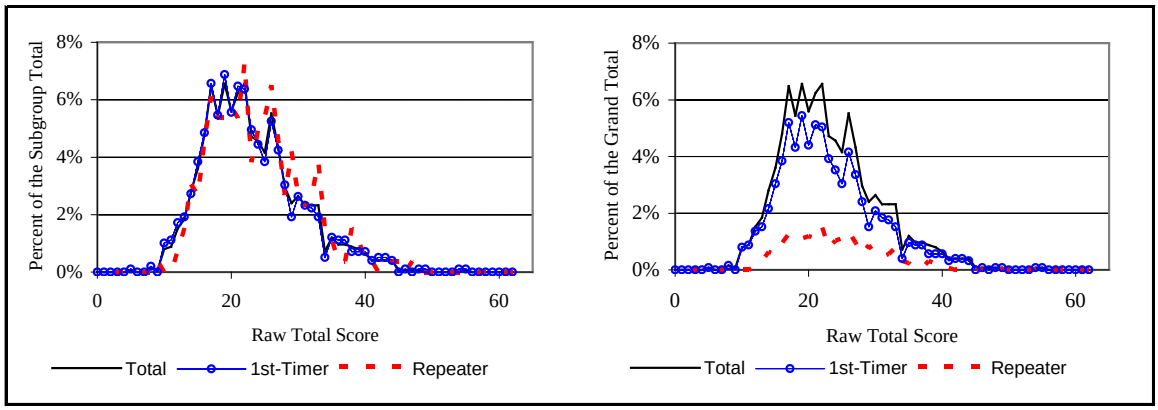

Figure 4. **Observed relative frequency distributions for Form B-Q.**

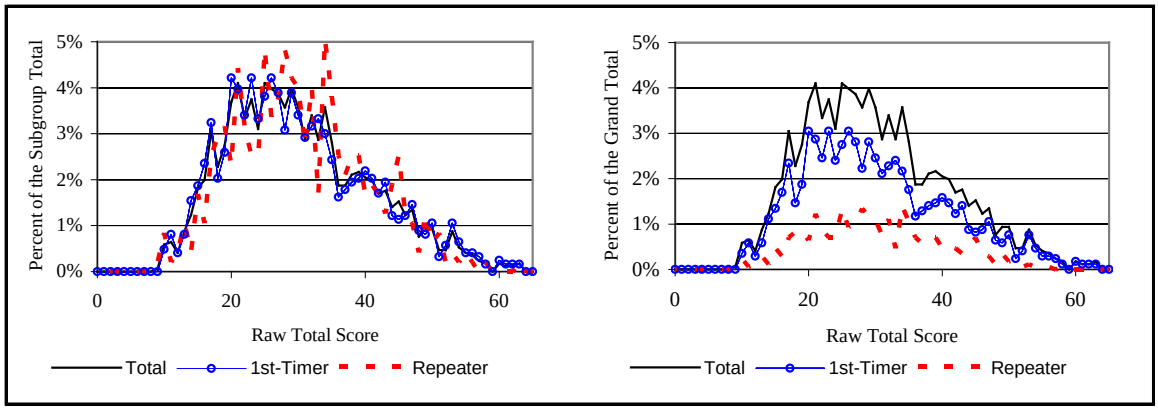

Figure 5. **Observed relative frequency distributions for Form C-Q.**

Overall, the figures show a significant degree of similarities between the repeater score distributions and the distributions for the first-time examinees, despite the more distinctive ups and downs in the repeater frequency distributions. Nevertheless, the location of the repeater distributions 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 do not look very pronounced on the plots). This suggests that, on average, the repeaters performed slightly better than the first-time examinees on the exam being studied.

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).

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 general repeater trends and data quality issues in mind, we considered equating results for various study forms. For each of the equating functions needed (for the total group, first-time examinee group, or 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 the data of a particular group. The evaluation and selection criteria were consistent with the criteria used for making operational equating decisions for the exam being studied. Table 7 summarizes the equating model selection outcomes for different groups on various forms. As shown in Table 7, the same equating model was chosen for all three study groups for each of the following new forms: A-V, A-Q, and C-Q. The smoothed chained

equipercentile model fit the data of various groups well for equating A-Q to RA-Q and for equating C-Q to RC-Q, whereas the equating relationship between A-V and RA-V appeared to be linear[5] for various groups.

Table 6

Significance of Group Mean Differences Between First-Timers and 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	< 0.0001
		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: $\mu_{repeater} - \mu_{first\text{-}timer} = 0$ for the two-tailed Z test.

Table 7

Equating Model Selected for Each Examinee Group for Various Forms

New form	Reference form	Examinee group	Equating model chosen
A-V	RA-V	1st-timer	Chained linear
		Repeater	Chained linear
		Total	Chained linear
A-Q	RA-Q	1st-timer	Smoothed chained equipercentile
		Repeater	Smoothed chained equipercentile
		Total	Smoothed chained equipercentile
B-V	RB-V	1st-timer	Smoothed chained equipercentile
		Repeater	Chained linear
		Total	Smoothed chained equipercentile
B-Q	RB-Q	1st-timer	Smoothed chained equipercentile
		Repeater	Chained linear
		Total	Smoothed chained equipercentile
C-Q	RC-Q	1st-timer	Smoothed chained equipercentile
		Repeater	Smoothed chained equipercentile
		Total	Smoothed chained equipercentile

As for new forms B-V and B-Q, the smoothed chained equipercentile model was appropriate for the total group and first-time examinee group equating, but the sparse data of the repeater group could not support the use of the chained equipercentile model. Therefore, the chained linear equating model was selected for the very small repeater group.

Repeater Effects on Score Equating

Evaluation outcomes of repeater effects on score equating is presented in this section. Results of average subpopulation invariance (i.e., the $RESD_j$ and *REMSD*) for various study equating were tabulated to show the overall invariance of the total-group scaled scores with respect to the repeater and first-time examinee subgroups. We also graphed the scaled score differences between equating outcomes and the *RMSD* outcomes with a band of ± 2 standard errors (for evaluating statistical significance) and a band for the DTMs (for evaluating practical

significance) to describe in details how the scaled scores resulting from equating based on different groups/subgroups data differed.

***RESD$_j$* and *REMSD* results.** Table 8 presents the *RESD$_j$* and *REMSD* results for each of the five new forms in this study. 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 between the total group and repeater subgroup in scaled scores.[6] Nevertheless, the scaled scores based on equating in the total group looked pretty invariant across subpopulations for various study forms overall, from a practical perspective.

In contrast to the practical significance outcomes, the statistical significance outcomes suggested a greater degree of subpopulation dependence problem for three of the five new forms—namely, A-Q, B-V, and B-Q. As shown in Table 8, several more of the $RESD_j$s and *REMSD*s were statistically significantly different from 0 (i.e., greater than +2 standard errors) than the two $RESD_j$s of practical significance. For forms A-Q and B-Q, the statistically significant subpopulation dependence problems were due to the equating differences between the total group and both of the subgroups. However, for form B-V, the problem was mainly due to the differences between the total group and the repeater subgroup.

Equating differences in scaled scores and *RMSD* results. To evaluate the invariance of scaled scores at individual new-form raw score levels, Figures 6 to 20 present the scaled score differences between equating based on data of various study groups for the five new forms in this study. And, Figures 21 to 25 present the *RMSD* results for the five new forms, respectively. These figures provide greater details about equating differences than the $RESD_j$s and *REMSD*s shown in Table 8. Figures 6 to 20 further allow the positive and negative scaled score differences to be observed. Specifically, Figures 6 to 10 show the scaled score differences for the first-time examinee group and repeater group (First-time examinees minus Repeaters). Figures 11 to 15 present the scaled score differences for the repeater group and total group (Repeaters minus Total group). And, Figures 16 to 20 present the scaled score differences for the first-time examinee group and total group (First-time examinees minus Total group). 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. Note that the scaled score differences in Figures 6, 11, and 16 were all linear because all of the equating functions for Form A-V were linear and the reference form scaling was also linear.

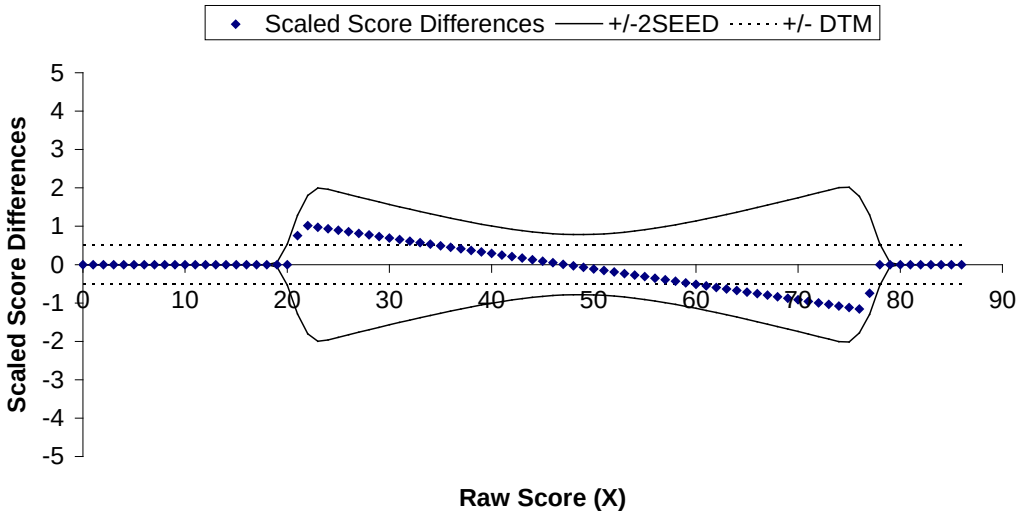

Figure 6. **Scaled score differences (First-Timers minus Repeaters) for A-V/RA-V.**

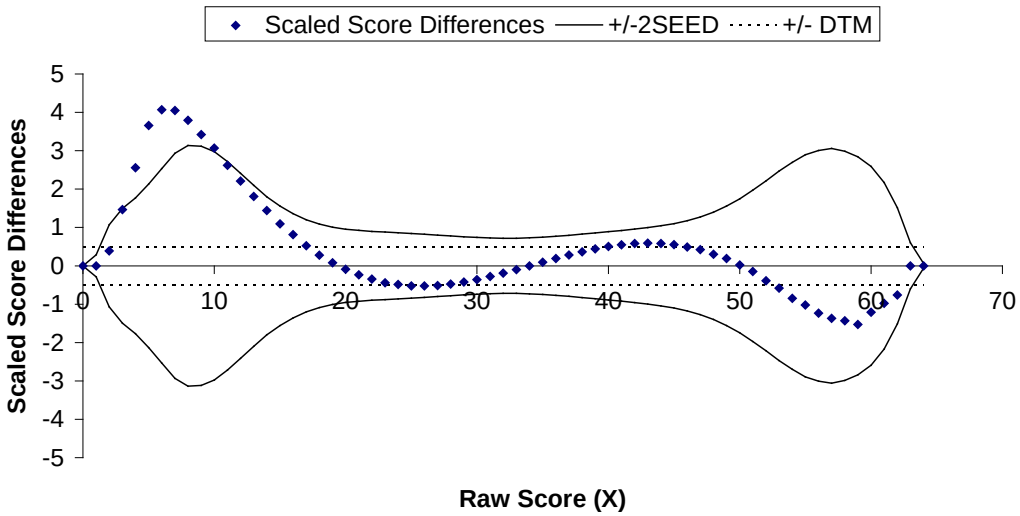

Figure 7. **Scaled score differences (First-Timers minus Repeaters) for A-Q/RA-Q.**

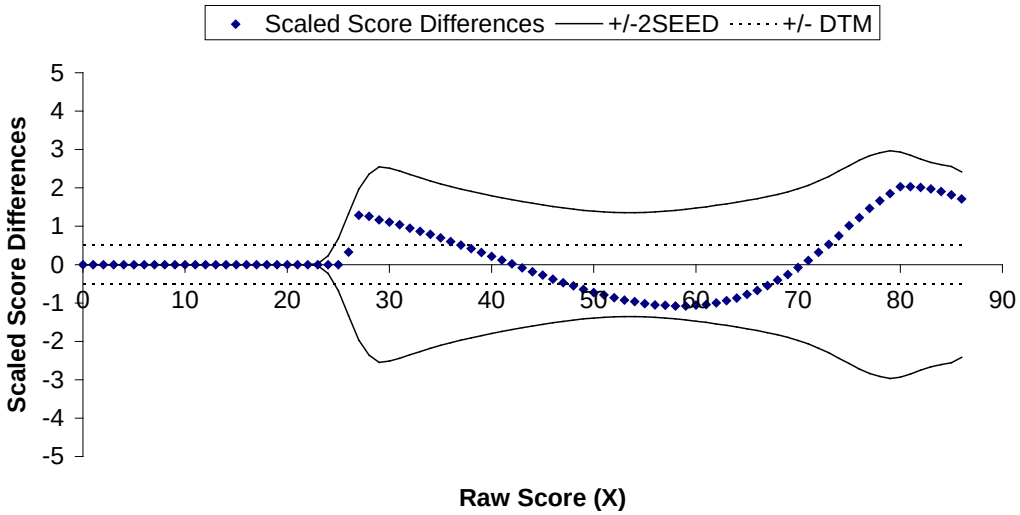

Figure 8. **Scaled score differences (First-Timers minus Repeaters) for B-V/RB-V.**

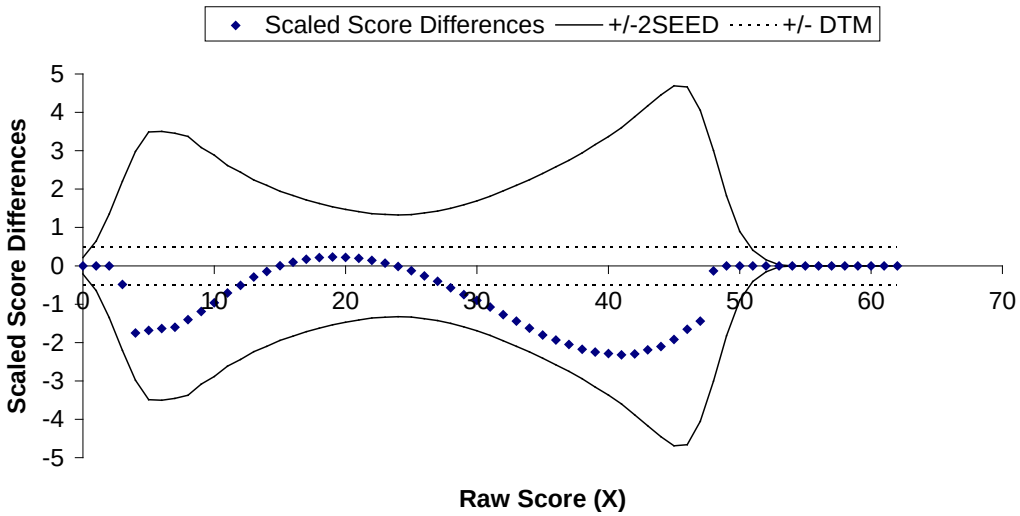

Figure 9. **Scaled score differences (First-Timers minus Repeaters) for B-Q/RB-Q.**

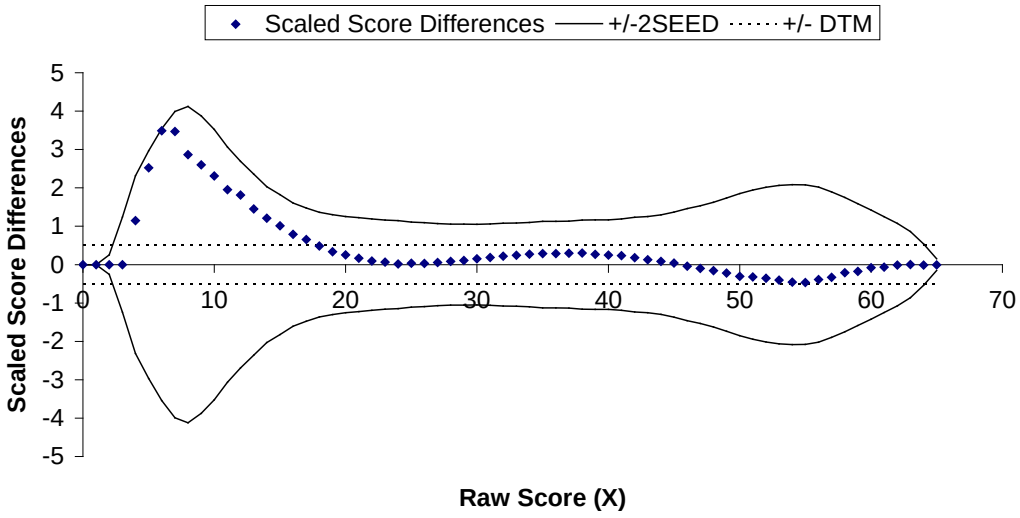

Figure 10. **Scaled score differences (First-Timers minus Repeaters) for C-Q/RC-Q.**

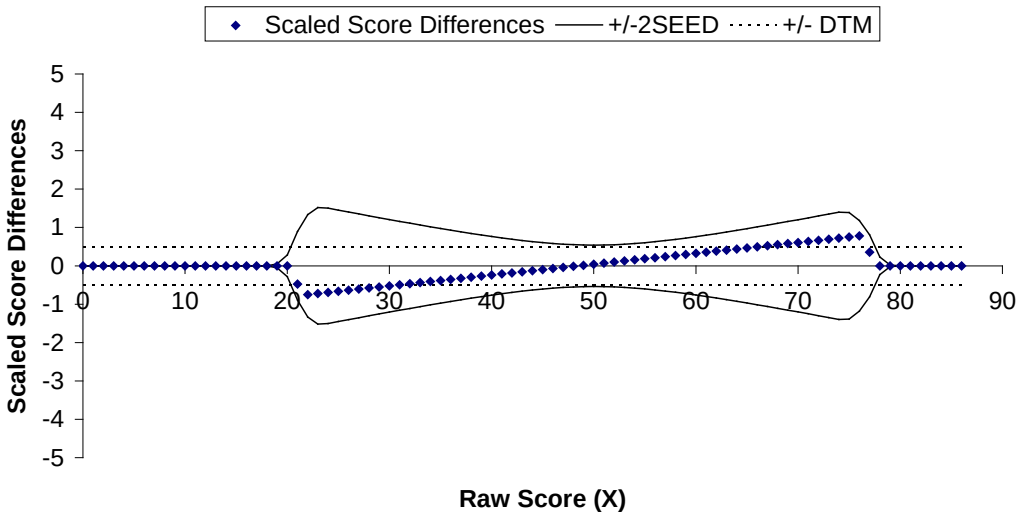

Figure 11. **Scaled score differences (Repeaters minus Total) for A-V/RA-V.**

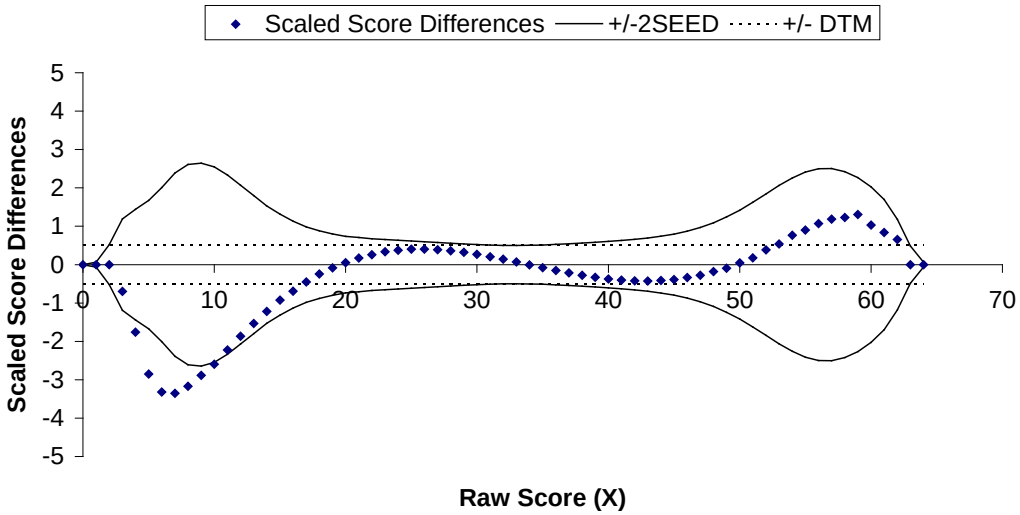

Figure 12. **Scaled score differences (Repeaters minus Total) for A-Q/RA-Q.**

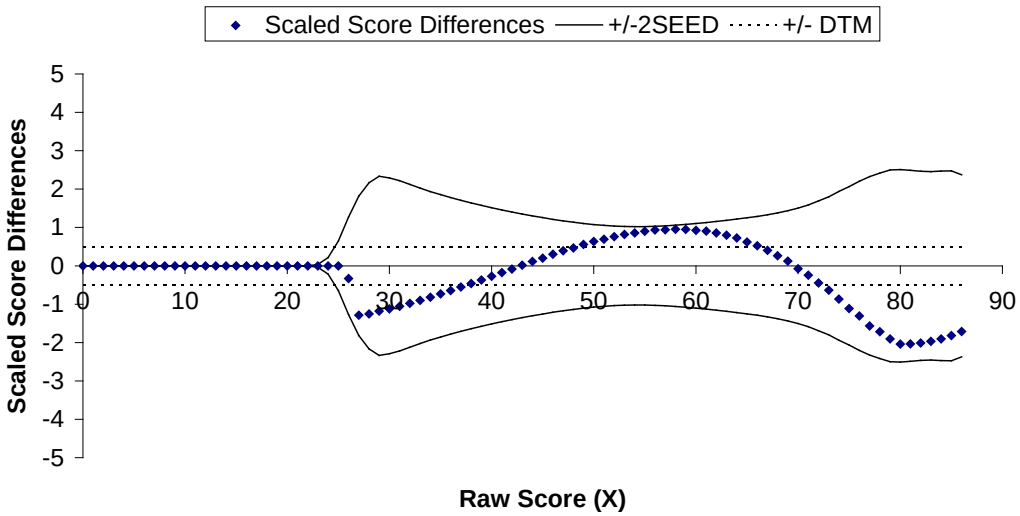

Figure 13. **Scaled score differences (Repeaters minus Total) for B-V/RB-V.**

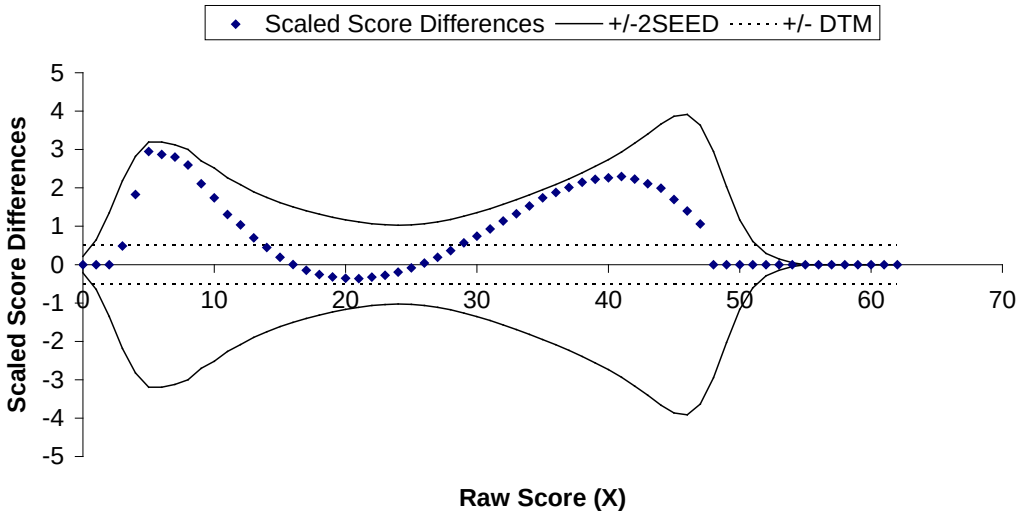

Figure 14. **Scaled score differences (Repeaters minus Total) for B-Q/RB-Q.**

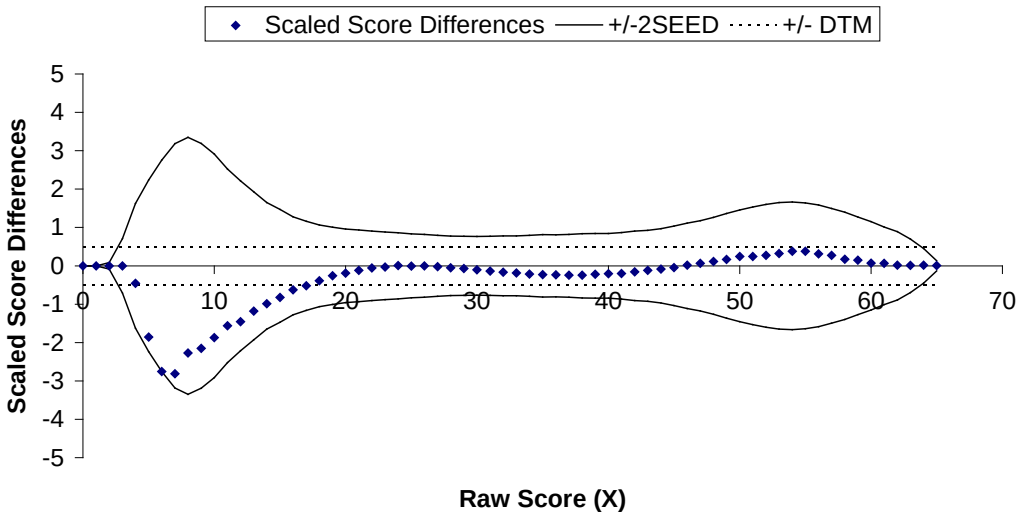

Figure 15. **Scaled score differences (Repeaters minus Total) for C-Q/RC-Q.**

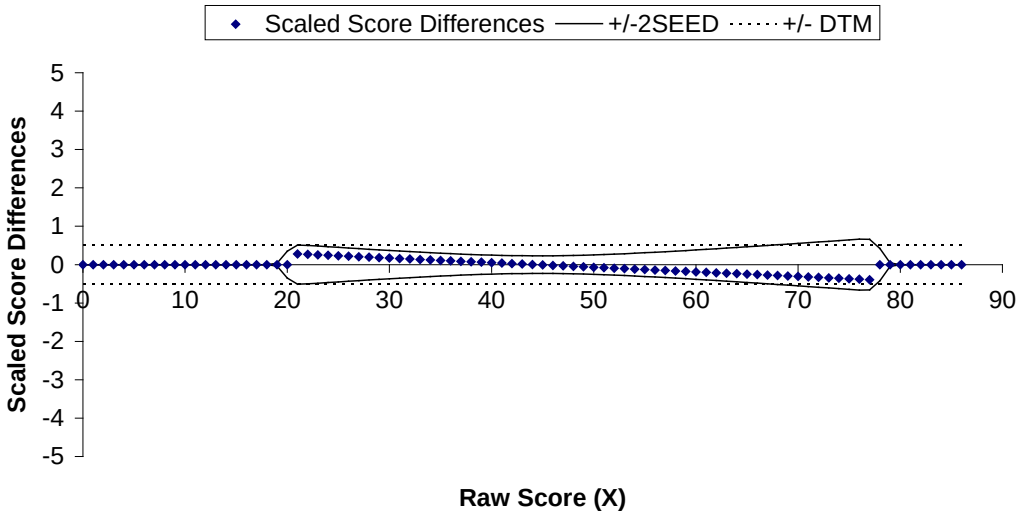

Figure 16. **Scaled score differences (First-Timers minus Total) for A-V/RA-V.**

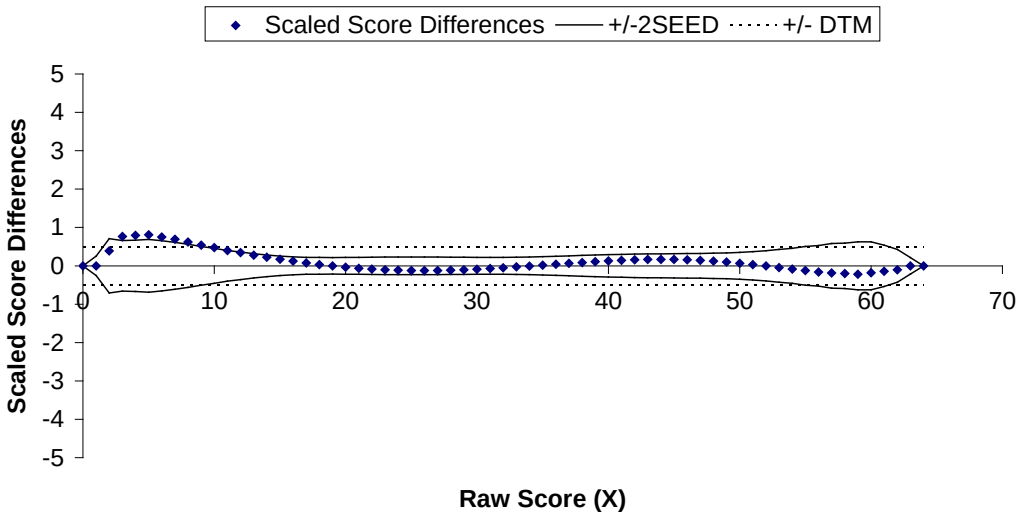

Figure 17. **Scaled score differences (First-Timers minus Total) for A-Q/RA-Q.**

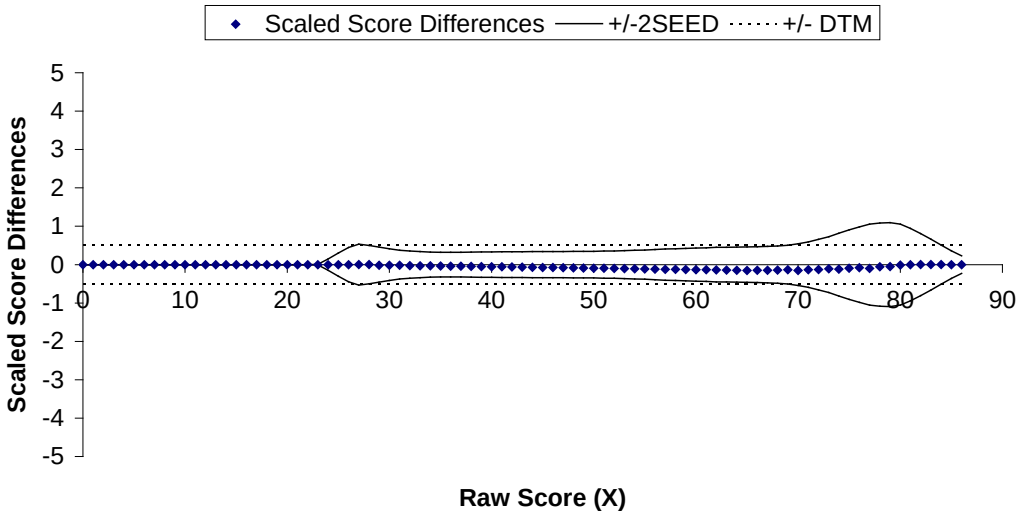

Figure 18. **Scaled score differences (First-Timers minus Total) for B-V/RB-V.**

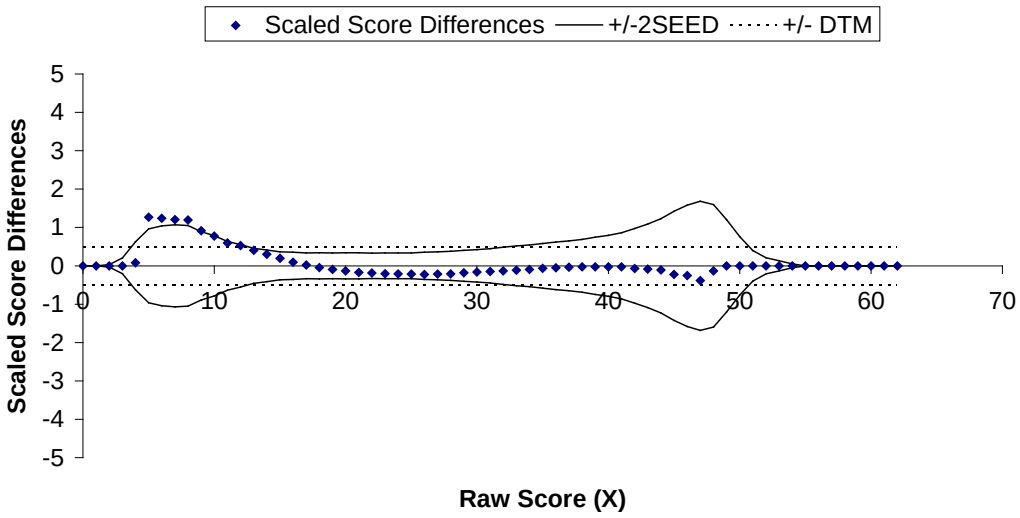

Figure 19. **Scaled score differences (First-Timers minus Total) for B-Q/RB-Q.**

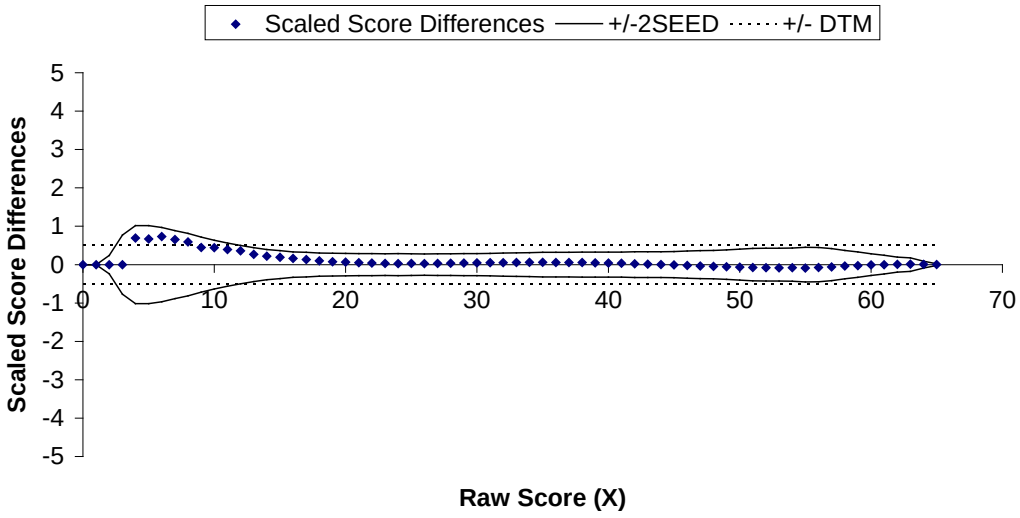

Figure 20. **Scaled score differences (First-Timers minus Total) for C-Q/RC-Q.**

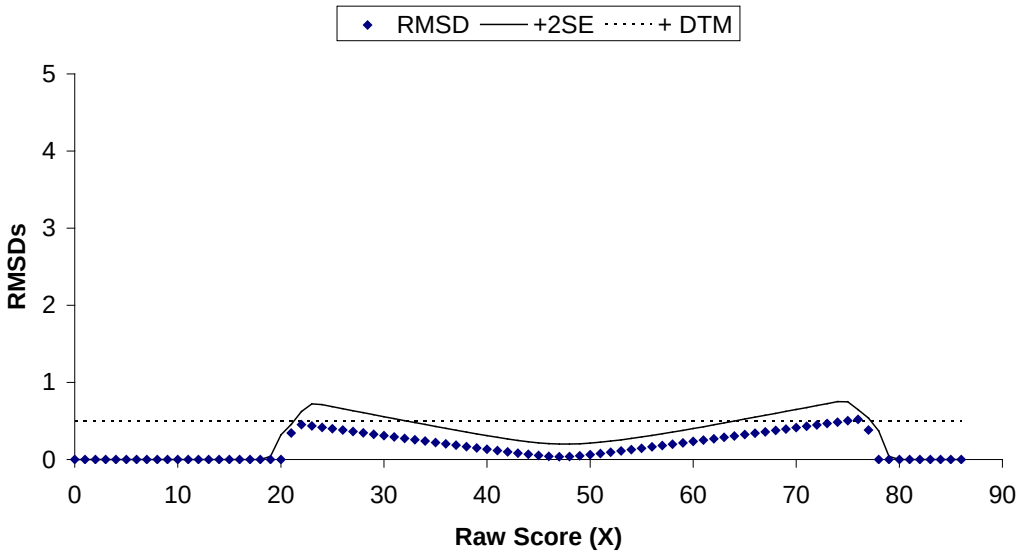

Figure 21. RMSD**s for A-V/RA-V.**

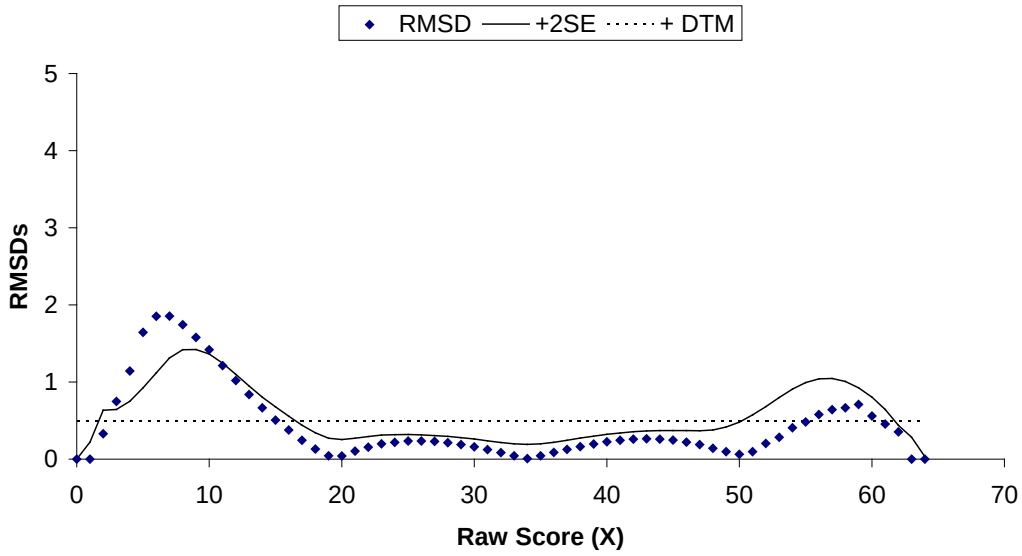

***Figure 22. RMSDs* for A-Q/RA-Q.**

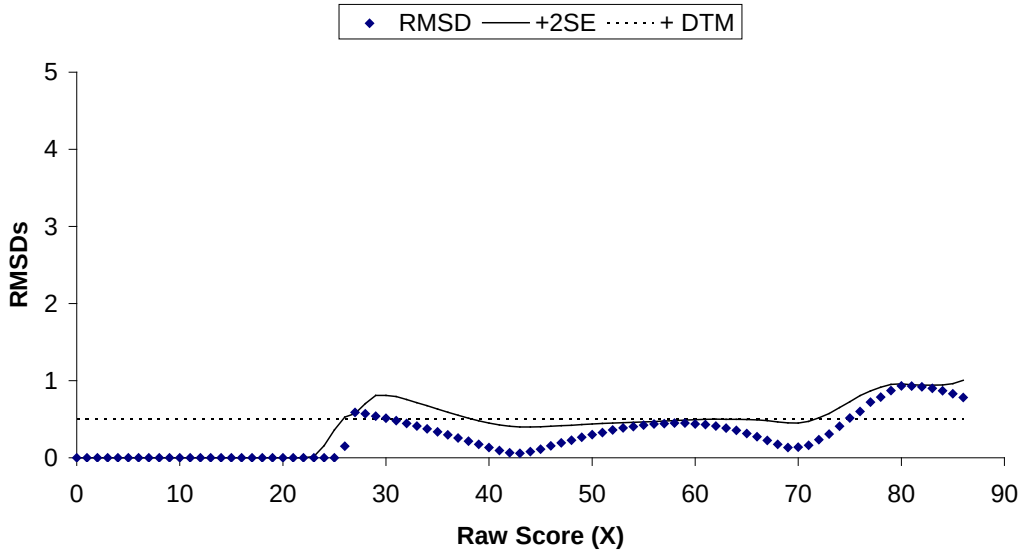

***Figure 23. RMSDs* for B-V/RB-V.**

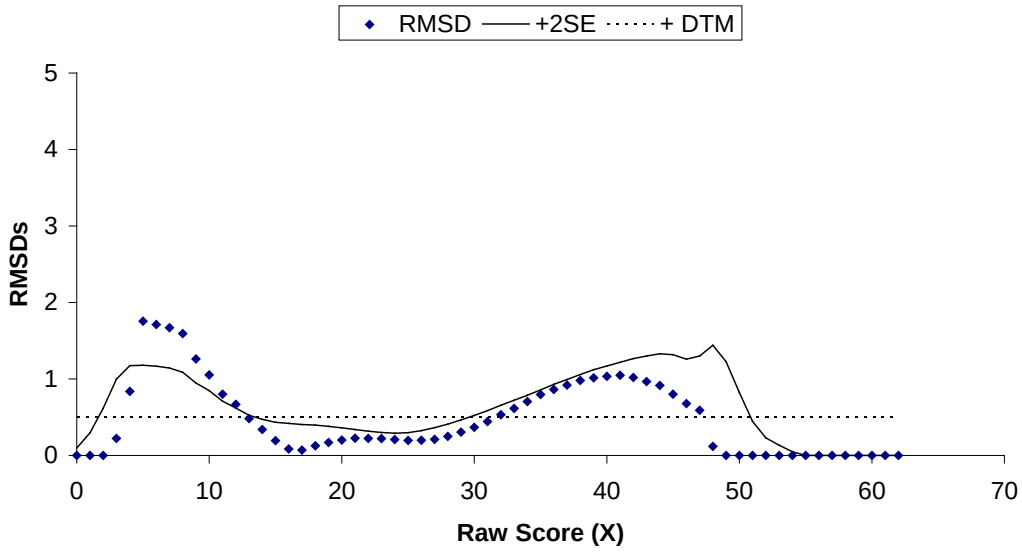

***Figure 24. RMSDs* for B-Q/RB-Q.**

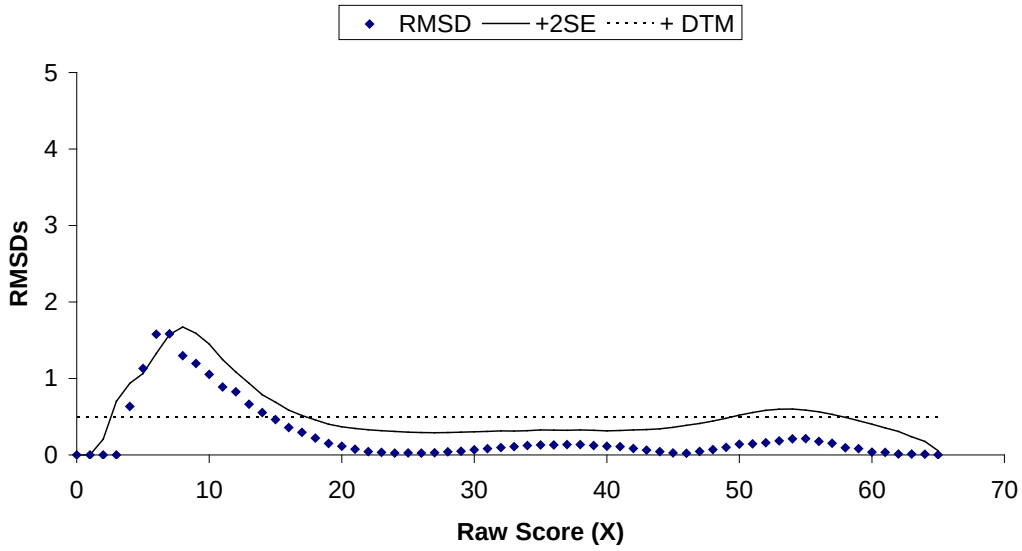

***Figure 25. RMSDs* for C-Q/RC-Q.**

Different than the *RESD$_j$s* and *REMSDs* in Table 8, the positive and negative equating differences in scaled scores exhibited in Figures 6 to 25 reveal that the scaled scores based on the repeater-group equating were not consistently higher or lower than those for the total-group or first-timer-group equating across study forms and measures. Practically or statistically significant scaled score differences usually occurred at the lowest and/or highest regions of the new-form raw score scale, especially for scaled score differences involving the repeater groups. Below, we present important findings based on Figures 6 to 25 in detail.

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

Despite that the scaled score differences shown in Figures 6 to 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.

Scaled score differences between repeater and total group equating. Figures 11 to 15 display the scaled score differences between equating in the repeater group and in the total group by new-form raw score levels. Similar to the findings based on Figures 6 to 10, we found many of the scaled score differences to be practically significant for each of the new forms, but most of the differences were not statistically significant. This is especially true for A-V, B-V, B-Q, and C-Q. The directions of equating differences shown in these figures also support the notion that the repeaters were at least as able as the first-time examinees on the study exam.

Scaled score differences between first-timer and total group equating. Figures 16 to 20 compare the scaled score outcomes based on equating in the first-timer group and in the total group. Such comparisons are commonly used to demonstrate effects of repeater performance on score equating, as the total group equating includes repeaters but the first-timer group equating excludes repeaters. Overall, Figures 16 to 20 show no statistically or practically significant differences in scaled scores between the total group and first-timer group equating for various new forms. This 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, in general, scaled score differences between the subgroups and total group equating were neither statistically nor practically significant in the middle region of the new-form raw score scale. However, for some forms there were significant scaled score differences at one or both ends of the scale. Specifically, for forms A-Q and B-Q there were many practically significant differences at both ends of the scale. While a large portion of the practically significant differences at the lower end were of statistical significance, those at the higher end were not. For C-Q, significant differences only occurred at the lower end of the scale, and most of the differences were practically but not statistically significant. For B-V, significant differences mainly clustered at the higher end, and only a few practically significant differences were borderline significant statistically. As to A-V, there were just a few practically significant differences at the higher end, and none of the differences were statistically significant. Overall, the *RMSD* results seem to imply potential (if not substantial) invariance problems for the total-group scaled scores for the low-achieving and high-achieving examinees.

Practical versus statistical significance outcomes. In sum, many of the scaled score differences shown in Figures 6 to 25 were practically significant, but not as many were statistically significant. This is especially true for Figures 6 to 15, which involved the repeaters and had rather large standard error bands around the scaled score differences. For Figures 16 to 20, though, the practical and statistical significance outcomes agreed with each other to a greater extent. This is because the scaled score differences between the first-timer and total groups were hardly of practical or statistical significance. And, while the *RMSD* results in Figures 21 to 25 showed consistent outcomes of practical and statistical insignificance for the middle region of the new-form raw score scale, on several study forms there were many inconsistent significance

outcomes at one end or both ends of the scale (specifically, some practically significant differences were not statistically significant).

The inconsistency between the practical and statistical significance outcomes seems to suggest an effect resulting from the measure used to assess significance. 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 of the practical evaluation criterion in the Discussion section.

A note on zero differences at the scale ends. Across study forms and various comparisons, some scaled score differences at the two ends of the raw score scale were perfectly zero. Most of the zero differences were consequences of truncating scaled scores that were out of the reporting scale range. Specifically, the out-of-range scaled scores were truncated to be the possible minimum or maximum of the reporting scale (i.e., scaled scores lower than 20 were set at 20, and scaled scores higher than 80 were set at 80). The reason for such a truncation was explained previously in the Method section (see subsection *A focus on raw-to-scale equating* and endnote 3). As out-of-range scaled scores were converted to have the same values, equating differences in the very low or very high scaled score regions were likely to become zero, leading to an impression of subpopulation invariance.

Comparisons of Various Invariance Measures

Some aspects of Figures 6 to 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 to 20), compared to the scaled score differences between the repeater group and total group (Figures 11 to 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 three new forms with statistically significant repeaters $RESD_j$s in Table 8 (A-Q, B-V, and B-Q), the scaled score differences exhibited in Figures 12 to 14 were practically significant at certain new-form raw score levels, but they were hardly significant statistically

because of the large standard error bands associated with the small repeater sample sizes. For the two new forms with statistically significant first-timers $RESD_j$s 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 most of the new-form raw score levels, except for those at the very low end of the scale (specifically, at and below the raw score level of 11 for A-Q, and at and below 13 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 new forms A-V, A-Q, and C-Q, whereas the $RESD_j$s and *REMSD* statistics in Table 8 suggested otherwise for these forms.

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

Highlight of Major Findings

- The self-reported repeater data used in this study looked reasonably sound. It was empirically verified and deemed to be the best option available for this study.
- Trends in repeater performance on the exam being studied—
    - On average, repeater scores across testing occasions looked fairly stable. Nevertheless, there were large scaled score gains/losses for a broad range of administrations. The test-retest correlation was 0.74 for Verbal and 0.72 for Quantitative for the overall repeater group.
    - High-performing examinees might not improve their scaled scores by retesting as the low-performing examinees could.

    - The repeaters were at least as able as the first-time examinees in general. Actually, the repeaters were more able than the first-time examinees on half of the study forms, but further studies are needed to better understand repeater performance patterns while taking into account demographic background information.

- Overall, the total-group equating function and its resulting scaled scores looked reasonably invariant across subpopulations. Differences in scaled scores between the total group and first-timer group equating were generally negligible from both practical and statistical perspectives. This implies that effects of repeater performance on score equating was not significant for the study exam.
- Although the repeater group equating seemed to be somewhat different from the equating in the first-timer group and total group, most of the scaled score differences that were of practical significance were not statistically significant. And, some of the results based on different invariance measures for the repeaters were mixed.
- Large equating differences on the reporting scale often occurred at the two ends of the new-form raw-score scale. Although the differences were often practically significant, they were seldom statistically significant.
- There might be an effect resulting from the measure used for evaluating the significance of equating differences in this study. The criterion for practical evaluation might be too sensitive in detecting significant scaled score differences, whereas the statistical evaluation criterion might be limited by small study sample sizes and, as a consequence, not able to reveal scaled score differences of importance.
- Overall, invariance outcomes based on various summary statistics seemed to be consistent to a reasonable degree. Discrepancies in the invariance outcomes were primarily associated with small repeater groups and could be reasonably explained.

Discussion

We elaborate on important study findings and their implications on test equating in this section. We will also discuss limitations of this study.

Effects of Repeater Performance

Overall, the insignificant differences between the total group and first-timer group equating imply negligible repeater effects on score equating for the exam being studied on both Verbal and Quantitative. Although equating in the repeater group looked different from equating for the first-timer and total groups, and the differences were sometimes of practical significance, most of the differences were not statistically significant (this is partly due to the small repeater sample size). Therefore, it seems safe to conclude that for the study exam the repeater effects on score equating were not significant. However, before rushing to replace the first-timer group equating with the total group equating to augment equating sample sizes and thus enhance equating precision, to be prudent one could wait until the repeater group equating outcomes are further scrutinized. Equating for the repeater group deserves to be further studied because of consistently observed equating differences that were practically but not statistically significant across exam administrations and measures. They seem to support the notion of unique repeater group equating. More research based on larger repeater group data is needed to investigate whether equating differences will remain statistically insignificant for the repeater group. And, it is important to reassess efficacy of the criteria used to evaluate practical significance of equating differences (to be further discussed below).

Practical Criteria for Evaluating Equating Differences

It is not uncommon to have conflicting statistical and practical significance evaluation outcomes. However, findings of this study seem to suggest that the practical evaluation criteria (i.e., the DTMs on the reporting scale) were too sensitive in detecting important equating differences, because most of the practically significant equating differences were not statistically significant. Although the statistically insignificant results could be partly explained by the small subgroup sample sizes, the pronounced disparities between the practical and statistical evaluation outcomes still cast doubt on the adequacy of the practical evaluation criteria. Despite that the practical criteria selected for this study looked reasonable in the context of cut-score decisions and were consistent with the operational rounding practice, they still had the disadvantages of being arbitrary and subjective.

Unrounded Versus Rounded Scaled Scores for Practical Evaluation

In addition to the selection of evaluation criteria, another important aspect of practical evaluation involved a decision on analysis score type (i.e., to use the unrounded or rounded scaled scores). At first, it seemed reasonable to use the rounded (integer) scaled scores to evaluate the practical significance of equating differences because rounded scaled scores were reported to candidates taking the exam being studied, which could be directly compared to some admissions screening threshold. However, we focused the comparisons of scaled score equating differences on unrounded values in this study to avoid potential confounding effects due to rounding, which could dramatically change the evaluation outcomes. If the rounded scores were used, some of the significant equating differences could be due to rounding instead of the equating.

Choice between unrounded and rounded scores primarily depends on whether the advantages outweigh the disadvantages. In an effort to assess the relative efficacy of these two score types, we compared equating differences based on rounded scores[7] to those based on unrounded scores using the same set of study data. Overall, patterns of equating differences based on the rounded and unrounded scores looked 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 large unrounded differences zero (e.g., both 40.49 and 39.50 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 result and the fact that in an invariance study we care more about scaled score differences due to equating than to rounding, it seems more important and appropriate to compare equating differences using unrounded scores rather than using rounded scores—even when the goal is to evaluate the "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 Self-Reported Repeater Data

In testing practice, often there is not a mechanism built in the registration and/or scoring system to automatically detect repeaters or effectively merge repeater score records across administrations. And, the post-administration analysis window is usually too narrow to allow

sufficient time for matching examinee records across administrations to manually identify repeaters. As a result, repeaters are often asked to identify themselves on a voluntary basis through a survey conducted at the registration or testing time. However, repeater information collected through examinees' voluntary responses is not likely to fully reflect the actual repeater phenomena. Reliability and validity of the repeater survey are often a cause for concern. This 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, in consideration of examinees' ability to recall their test-taking history and their willingness to respond to a lengthy survey, design of the repeater survey often falls short in acquiring sufficient information for analysis purposes.

Although we used the self-reported repeater data to facilitate our analyses in this study, we strived to verify the repeater information using empirical examinee records. Given the availability and limitations of our study data, it was the best option we could have. The verification process was labor-intensive and time-consuming, but it helped to raise our confidence in the self-reported repeater data. If we had used the empirically matched examinee records for our analyses instead, we would have under-identified repeaters on the study exam because the aggregation of empirical records across administrations was restricted by a lack of effective matching variables. Nevertheless, we would have more confidence in the resulting study findings if we could have a better way to identify repeaters or to more thoroughly verify the self-reported data. 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.

Impact of Verbal Versus Quantitative Invariance Outcomes

The effects of repeater performance on score equating for Verbal and Quantitative were quite similar based on the findings of this study. However, because operationally the Verbal subscore is weighted more (by 10%) than the Quantitative subscore in calculating the composite score, the subpopulation invariance properties of the Verbal equating would have a stronger impact on the scaled composite score outcomes than that for 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 Versus Specific Repeater Effects

Partly because of the limited sample size for study, we focused on the overall, nonspecific repeater effects, instead of the effects of specific repeater subgroups. We consider the investigation of such overall, nonspecific effects an important step toward unveiling the potential impact of repeater performance on equating and scoring, and the in-depth examination of general repeater patterns would shed some light for future research. Nonetheless, to better untangle sources of repeater effects, more research that focuses on specifically defined repeater subgroups differing in the number of retakes, test-retest time interval, and so on, is needed.

In this study, we did not consider atypical cases when examinees repeated the same test or anchor test. Based on our experience, the frequency of such cases was very low for the exam being studied. To avoid any security problems or fairness issues, one should strive to prevent such cases from happening even before equating takes place, anyway.

Range Restriction Due to Self-Selected Repeaters

The repeater population is usually restricted in range because of the self-selected 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 cases, 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 average 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 testing program. While considering the effects of range restriction on score equating, we first need to have a good understanding of the range restriction issues pertinent to a particular testing program.

Limitation Due to Reference-to-Scale Conversion

Aside from the aforementioned study limitations, we also faced a trade-off between equating practicality and precision. In this study, we decided to focus on the raw-to-scale conversions primarily because of their practical importance to score fairness. It was also because

repeater effects on equating were seldom studied in the context of scaled score conversions.[8] However, the new form raw-to-scale conversion function depends on the previously established reference-to-scale conversion function, which might be subject to effects of repeater performance unless the effects were controlled for in previous equating. As a result, as much as the repeater effects were controlled for in the raw-to-raw equating of this study, study outcomes based on the raw-to-scale conversion could not be free of potential bias due to the involvement of the reference-to-scale conversion. In essence, scaled score differences in this study reflected not only the differences in raw-to-raw equating but also the characteristics of the reference-to-scale conversions, as well as the treatment of equating/scaling outcomes (e.g., rounding and truncation) for score reporting purposes.

It is important to study effects of repeater performance on score equating, and the study outcomes should guide the design and shape the strategies for future equating. As characteristics of repeater performance are intricately related to equating outcomes, we need to compliment equating studies with information about repeater performance trends. In addition, more research based on data featuring varying repeater characteristics from a variety of testing programs is needed to broaden the range of evidence of repeater effects on equating.

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.

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, NY.

Harris, D. J. (1993). *Practical issues in equating*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Kingston, N., & Turner, N. (1984). *Analysis of score change patterns of examinees repeating the Graduate Record Examinations General Test* (Research Report No. RR-84-22). Princeton, NJ: ETS.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York, NY: 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* (Research Report No. RR-06-20). Princeton, NJ: ETS.

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.

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* (Research Memorandum No. RM-08-05). Princeton, NJ: ETS.

Notes

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

[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.

[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.

[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.

[5] There was not much curvilinearity in the equating relationship between forms A-V and RA-V, based on the data of various study groups, especially where there were substantial amount of data. And, driven by scarce data at the high end of the distribution, the smoothed chained equipercentile equating line went unreasonably high at the top. Between the Tucker and chained linear lines, the chained linear line was more consistent with the data.

[6] However, this practical significance of equating differences could also be a function of the use of smoothed chained equipercentile equating model for the total group but chained linear model for the repeater subgroup for Forms B-V and B-Q. As previously described in the Method section, this potential equating method effect is a trade-off for maintaining equating model fit for various sets of study data.

[7] The numerical values of the practical significance criteria 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.

[8] On top of the raw-to-raw equating, raw-to-scale conversions require additional computations, transformations, and/or truncations, which all add to the complexity of scaled score outcomes. As a result, raw-to-scale conversions are less frequently studied in equating research than raw-to-raw conversions.