Effect of Repeaters on Score Equating in a Large Scale Licensure Test

Sooyeon Kim
Michael E. Walker
ETS, Princeton, NJ

Paper presented at the annual meeting of the
American Educational Research Association (AERA) and the
National Council on Measurement in Education (NCME)

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

Unpublished Work Copyright © 2009 by Educational Testing Service.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unpublished, proprietary work of ETS.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

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

This study investigated the subgroup invariance of equating functions for a licensure test in the context of a non-equivalent groups with anchor test (NEAT) design. Examinees who had taken a new, to-be-equated form of the test were divided into three subgroups according to their previous testing experience: (1) repeaters who previously took the reference form, to which the new form would be equated; (2) repeaters who previously took any form other than the reference form; and (3) first-time test takers for whom the new form was the first exposure to the test. Equating functions obtained with two subgroups, all repeaters and first-time test takers, were similar to those obtained with the total group, supporting score equatability of the two forms. When the repeater subgroup was divided into two groups based upon the particular form examinees took previously, however, subgroup equating functions differed substantially from the total group equating function, indicating subgroup dependency of score equating. The results indicate that repeater membership needs to be more clearly specified to assess the impact of repeaters on score equating. Such clarification may be particularly necessary for high-stakes licensure tests, on which repeaters tend to perform more poorly than do first-time test takers.
Introduction

In most if not all high-stakes testing programs, different forms of a single test are used in different administrations to help prevent unfair advantages for those who may have access to information from previous administrations. Although the different test forms are designed to measure the same content and to have the same statistical characteristics, they often differ slightly from one another. To ensure fairness to examinees taking different forms, the scores on the forms are adjusted to make them equivalent to one another.

Test equating is a statistical method for adjusting for differences in difficulty among forms built to the same specifications. One of the most basic requirements for score equating is that equating functions should be subpopulation invariant (Dorans & Holland, 2000); that is, the equating function must operate independently of subgroups of examinees from whom the data are drawn to develop the conversions. Lack of subgroup invariance in an equating function indicates that the differential difficulty of the two forms is not consistent across those subgroups. Accordingly, the new and reference test forms are not equatable and their interchangeability is questionable.

This study investigated the population invariance of score equating by comparing equating functions derived using two distinct subgroups from the examinee population for a large-scale licensure test. The two subgroups were repeaters (those who had taken the test before) and first-time test takers (those taking the test for the first time, hereafter called first-timers). Comparing equating functions for repeaters and first-timers is particularly important for licensure tests because, on such tests, these groups differ
considerably in ability. In general, examinees’ need to retake the test indicates that they did not achieve a score needed to obtain licensure on the first attempt.

The few studies that have examined the impact of repeating examinees on equating showed inconsistent results (Andrulis, Starr, & Furst, 1978; Cope, 1985; Puhan, 2008). Andrulis et al. (1978) compared an equating function derived from the total sample to the function derived from a subsample of first-timers. They found that the performance of repeaters tended to move the raw scores that map to the passing scores downward; this effect became more pronounced as the number of repeaters increased. On the basis of this result, Andrulis et al. suggested that repeaters should be excluded from the equating conducted to produce the conversion for reporting scores; this conversion then could be applied to the total sample including both repeaters and first-timers. Cope (1985) obtained somewhat different results; equating functions derived using only first-timers did not differ from those derived using the total sample. On the basis of this result, Cope concentrated on the practice of excluding repeaters from the equating sample and suggested further research to assess the impact of eliminating large numbers of repeaters on the resulting equating function. Recently, Puhan (2008) examined the impact of including repeaters in the equating sample on final equating results using data from certification tests. The differences between the equatings derived using all examinees or only first-timers were very small, with no practical impact on examinees’ pass/fail designations.

Excluding repeaters from the equating sample may be reasonable under some circumstances; however, it might create an equating group that is not representative of the group that was tested, especially if many examinees repeated the test. It could also lead to
other issues, such as an increase the standard error of equating because of the reduction in the sample size (Kolen & Brennan, 2004). Inclusion of repeaters, however, may produce other concerns if those repeaters in the new form sample performed substantially better on the common (i.e., anchor) items than on the remainder of the test because of previous exposure. In such a case, the new form sample would appear to be more proficient than the reference form sample; and as a consequence, the new form would appear more difficult than it really was. This incorrect adjustment could be substantial if the proportion of repeaters were large.

Perhaps the most commonly used equating design is the non-equivalent groups with anchor test (NEAT) design, in which each test form is administered to one of two groups that differ in ability. The major motivation for considering repeaters under the NEAT design is that they may tend to perform better on the common anchor items than on the unique new items. We would expect this argument to hold if repeaters were previously exposed to the reference form to which the new form would be equated. At the same time, we would expect repeaters who took any form other than the reference form to perform similarly on both common and unique items, as both types of items would be new to these examinees.\(^1\) If so, repeaters should be divided into two groups, producing three subgroups of examinees: (1) first-timers, who never took the test before; (2) repeaters who took the reference form, hereafter called repeaters (reference), and (3)

\(^1\) It is possible that the test items, once exposed, could be compromised (e.g., by being posted on the internet or otherwise widely distributed). In this case, the anchor and unique items would behave differently even for repeaters who had not previously taken the anchor items. Note that in this case, the same would be true for first-timers, so that no distinction could be made between these two groups. Thus, we do not consider this case here.
repeaters who took any form other than the reference form, hereafter called repeaters (other forms).

A subgroup invariance study, which deals with examinees’ repeater status as a major subgroup factor, offers an increased understanding of the impact of sample selection on test score equating. The present study examined subgroup dependency of the equating function using two classifications of examinees into subgroups based on repeater status. The first classification included two subgroups, (1) first-timers and (2) undifferentiated repeaters. The second classification included three subgroups, (1) first-timers, (2) repeaters (reference), and (3) repeaters (other forms). The impact of the repeater effect was assessed by comparing the equating functions derived on the different subgroups.

Method

Data

Data sets from two national administrations of a large-scale licensure test were used. The data were collected using a NEAT design. The test consisted of 107 multiple-choice items; 42 items were common across new form $X$ and reference form $Y$. Descriptive statistics are summarized in Table 1. The table shows that the total new form group was as adept as the total reference form group. Their mean scores on the anchor differed by only 0.14 correct answers, an effect size of 0.03, indicating a negligible difference between the two populations. The difference in the mean scores on the two forms (an effect size of -0.36) therefore appears to be attributable to the difference in difficulty between the two forms. In both the new form and reference form groups, the first-timer group was higher in ability as measured by the raw test scores than the
repeater group, yielding a mean difference larger than one standard deviation. Figure 1 presents the relative frequency distributions of new form scores in the three subgroups, given as percentages of the total group.

Procedure

Figure 2 presents the general framework of possible equatings using different subgroups. In the NEAT design, the equating functions derived using either repeaters or first-timers were compared to the equating function derived using all examinees to determine whether the resulting equating function from the total group would produce comparable scores to the subgroup equating functions regardless of the examinees’ previous testing experience. The present study included the following three steps.

Step 1: Obtain total group and subgroup equating functions. The equating relationship between the new \((X)\) and reference \((Y)\) forms was derived using three groups, (1) total examinees, (2) first-timers, and (3) repeaters. These equating relationships are represented with solid lines in Figure 2. Two additional equating functions were derived using the two subgroups of repeaters, (1) repeaters who took the reference form and (2) repeaters who took any form other than the reference form. These equating relationships are represented with dashed lines in Figure 2.

Step 2: Compare total group and subgroup equating functions. The equating function derived using each subgroup configuration was compared to the total group equating function. The differences were quantified across all subgroups using the Root-Mean-Square Difference (RMSD) and Root-Expected-Mean-Square Difference (REMSD) deviance measures (Equations 1 and 2 to follow). In general, a negligible difference indicates that the equating relationship is not influenced significantly by the
subgroups used in deriving that function. In addition, the difference between subgroup equating functions and the total group equating function was separately quantified for each subgroup to assess more clearly the impact of reporting scores for that subgroup based on the total group equating transformation (Root-Expected-Square Difference, or RESD, in Equation 4 to follow). The impact of three different conversions on examinees’ pass/fail designations was assessed.

Step 3: Estimate precision of the deviance measures. A total of 2,000 replications were obtained in each equating design using a bootstrap resampling technique (as implemented in SAS PROC SURVEYSELECT) to estimate a 95% confidence band for the RMSD measure conditioned on each raw score. In each replication, examinees were randomly drawn with replacement from each reference and new form group until bootstrap samples included the same number of examinees as did the actual reference and new form groups. Both the new and reference form samples then were divided into either two or three mutually exclusive subgroups according to examinees’ repeater membership. In each replication, three groups (total group, repeater subgroup, first-timer subgroup) were formed in the two-subgroup condition. The repeater subgroup was divided into repeater (reference) and repeater (other forms) in the three-subgroup condition. The new form scores were equated to the reference form for the total group and the two or three subgroups in each replication using the chained equipercentile method, and the RMSD was calculated using both equal and proportional (i.e., unequal) weights. For the two weighting methods, which method produced the larger RMSD values depended on how the difference in equated scores was paired with the group weight for the subgroups. The
95% confidence interval (CI) for the RMSD measure, which covers RMSDs in the 2.5\textsuperscript{th} to 97.5\textsuperscript{th} percentile range, was constructed on the basis of the 2,000 replications.

**Equating Method**

The chained equipercentile equating method was used to derive the equating relationship between the new and reference forms of the test. The data were pre-smoothed using a log-linear model that preserved the first five univariate moments of each marginal distribution (i.e., of the total score and of the anchor score). No bivariate moments were preserved.\(^2\)

**Deviance Measures**

Four deviance indices were used as population invariance measures and calculated using the following formulas. The RMSD was defined in von Davier, Holland, and Thayer (2004) for the anchor test or NEAT data collection design, and the REMSD was defined in Holland (2003). The REMSD index was used to obtain a single value summarizing the values of RMSD\((x)\) over the distribution of \(x\) in the total group. The ewREMSD (Kolen & Brennan, 2004, p. 443), which gives equal weight to all score points - the inverse of the total number of score points (.1 in this case), was also calculated, particularly for the cut-score region (63 to 72)\(^3\) to examine the impact of

\(^2\) Preserving only univariate moments and no bivariate moment in this situation results in a slightly better fit of the marginal distributions than when the first bivariate moment is also preserved. Such a strategy is possible here because chained equating operates only on the margins. In any event, differences in equating results with and without preserving the bivariate moment are negligible.

\(^3\) All states do not use the same cut score.
subgroup influence on the examinees’ pass/fail designations. In addition, the RESD was computed as the weighted average of the squared differences between each subgroup equating function and the total group equating function at each raw score level. Thus, each subgroup would have a single summary RESD value. Note that we did not standardize any of the measures in this study by dividing by the standard deviation, as is the usual practice. We could use unstandardized measures because we did not conduct any comparisons across different testing programs. Using raw measures left the deviance indices in the metric of the raw scores, which facilitated interpretation of results (see Liu, Feigenbaum, & Dorans, 2005, for another example of this practice). The various indices are defined as follows.

\[
RMSD(x) = \sqrt{ \sum_j w_j [e_{yji}(x) - e_{yi}(x)]^2 }, \quad (1)
\]

\[
REMSD = \sqrt{ \sum_j w_j \sum_i r_i [e_{yij}(x) - e_{yij}(x)]^2 }, \quad (2)
\]

\[
ewREMSD = \sqrt{ \sum_j w_j \sum_i [e_{yij}(x) - e_{yij}(x)]^2 }, \quad (3)
\]

\[
RESD_j = \sqrt{ \sum_i r_i [e_{yij}(x) - e_{yij}(x)]^2 }, \quad (4)
\]

where \( x \) represents each raw score point, \( e_{yji}(x) \) indicates the equating function in the \( j \)th subgroup, \( e_{yi}(x) \) represents the equating function in the total group, \( w_j \) denotes the proportion of subgroup \( j \) in the total group, and \( r_i \) indicates the relative proportion of examinees in the total group at each raw score level.

As shown in Table 1, the subgroup sizes were substantially unbalanced. The proportion of first-timers was over 60%, whereas the proportion of repeaters (reference)
was lower than 10% in both groups. Thus, both the RMSD and REMSD measures would be heavily influenced by the first-timer subgroup. The RMSD measure itself might be biased against smaller groups (Yang, 2004). As a simple way to adjust for the unbalanced proportions, equal weight was imposed on each group when calculating RMSD and REMSD (e.g., \( w_j \) was always .5 in the two subgroup condition, \( w_j \) was always .333 in the three subgroup condition), along with the proportional weight derived from the actual relative size of each subgroup.

To determine when the REMSD was large enough to justify concern about the equatability of two forms, the notion of the score “Difference That Matters” (DTM: Dorans & Feigenbaum, 1994), defined as half of a raw score point in the raw-to-raw score transformations (i.e., 0.5), was used. Furthermore, the 95% CI of the RMSD was generated using a bootstrap resampling technique to evaluate equating difference from a statistical perspective.

Results

Preliminary Analysis

To clarify the effect of previous exposure to common anchor items on the part of the repeater (reference) group, moderated regression analyses were performed predicting new item or non-anchor scores from anchor item scores, repeater membership (reference vs. other forms), and the anchor × repeater membership interaction. From this overall analysis, separate regression lines for the two repeater subgroups then were computed to predict new item scores from the anchor score, by plugging repeater membership
information (e.g., 1 for repeater–reference and 0 for repeater–other forms) into the moderated regression equation.

Figures 3 and 4 present the separate regression lines for the new form group and the reference group, respectively. As shown, the pattern of regression lines and statistical results for their parameters were very similar in both test form groups. The intercepts for the two repeater subgroups were statistically significantly different but the slopes were not. The pattern of results indicates that equal performance on the anchor test predicts lower overall performance on the non-anchor items for those who have previously seen the anchor than for those who have not. In other words, the repeater (reference) group performs better on the anchor without a commensurate improvement in performance on other items. This finding is important insofar as it indicates the necessity of differentiating repeaters based upon the particular form they took previously.

Subgroup Equating Analysis

Figure 5 presents the chained equipercentile equating functions, derived using the total group and various subgroups, at the new-form raw score points where most examinees (about 98%) in the total group were located. The equating functions derived from repeaters (total), repeaters (other forms), and first-timers were very similar to the equating function derived from the total, indicating negligible differences among them. The equating function derived using the repeater (reference) subgroup, however, differed from those derived using the other groups, leading to differences in equated raw scores.

Figure 6 depicts equated raw score differences between each subgroup equating and total group equating, along with the DTM criterion denoted by dashed lines. The solid line at zero denotes the total group equating. The differences between the total and
subgroup equatings fell within .5 raw score units for the repeater (total) and repeater (other forms) subgroups. The difference for the first-timer subgroup was between the DTM band for raw scores higher than 57, but fell outside the DTM range for raw scores lower than 57. This can be explained by the fact that the total equating function, particularly for the low score region, was determined mainly by the repeaters, who tend to perform poorly, rather than the first-timers (see Figure 1).

As expected, the equating function derived using the repeater (reference) group was substantially different from the equating function derived using the total groups, yielding large positive differences for all the score points. The direction indicates that the new form was treated as much more difficult than the reference form in this condition as compared with the total group equating. Because repeaters (reference) performed better on common or anchor items than on new or non-anchor items due to either item exposure or practice, the repeaters appeared more able than they would have had they not seen the common items previously. Because repeaters (reference) did not do as well on the new or non-anchor items, the new form appeared more difficult in the process of score equating. The differences were extremely large for raw scores higher than 80. As displayed in Figure 1, however, almost no repeaters (reference) attained those high raw scores. These extreme differences were therefore somewhat artificial, caused by a lack of data.

Figure 7 depicts how RMSD, shown along with the DTM and REMSD measures, varies across score levels for the new form when the total group is divided into two subgroups, repeaters and first-timers. Both equal weights (e.g., .5 for each group) and unequal weights (proportional to subgroup size) were applied to the two subgroups to
calculate the RMSD measure, but the difference between them was negligible across the score range. The solid line is the REMSD value of .24, which was almost the same regardless of the weighting methods. The dotted line at .5 denotes the DTM. The REMSD measure was about one-half of the DTM, indicating that the equating function for each subgroup was sufficiently close to that of the total group. The RMSD values were much smaller than the DTM for most raw score points, including the cut-score range (raw scores of 63 to 72). The RMSD was larger than the DTM, however, at the lowest and highest scores. Given the closeness of all three equating functions as presented in Figure 6, and the practically negligible RMSD and REMSD values for most raw score points, the resulting equating function for this form can be considered invariant for both subgroups. In addition, the RESD values, which were separately calculated for each subgroup to assess subgroup dependence, were .20 for the repeater subgroup and .27 for the first-timer subgroup; they also were practically negligible. The ewREMSD values (.11), summarized over the cut-score region, were almost negligible under the two subgroup condition.

Figure 8 shows the same comparison as Figure 7, but with repeaters classified into those who took the reference form previously and those who took other forms. With the repeaters thus differentiated, the patterns of RMSD and REMSD changed dramatically, and the impact of weights was clearly noticeable. When the subgroups were weighted by their actual proportions in the total groups, the REMSD summary measure was twice as large as the DTM, and the RMSD values were larger than the DTM at the low and high ends of the scale. The RMSD values were smaller than the DTM, however, for raw scores in the range 62 to 74, which includes the cut-score range of 63 to 72. This trend was more
salient when the three subgroups were considered equally important regardless of their actual proportion in the total group. The RMSD values were much larger than the DTM for the entire score range, and the REMSD value (2.22) was four times as large as DTM, clearly indicating subgroup dependence of equating functions. The extremely large RMSD values for the upper raw score range, however, is likely to be an artifact of the lack of data (see Figure 1). Under the three subgroup condition, the RESD values were 3.81 for repeaters (reference) and .41 for repeaters (other forms). The $ew$REMSD value (.86), summarized over the cut-score region, was much larger than DTM under the condition where the three subgroups were equally weighted. Again, such large RESD and $ew$REMSD values for repeaters (reference) indicate a substantial linking difference between the subgroup and the total group, which will lead to different reported scores for the same examinees depending on the groups used for equating.

The results presented above were based on a single data set. The bootstrap resampling technique was applied to assess the variability of the RMSD values across 2,000 samples. The 95% CIs for the RMSD values are depicted in Figures 9 to 12. Figure 9 presents the 95% CI of RMSD with undifferentiated repeater membership and both repeaters and first-timers weighted by their proportions in the total group. The 95% CI fell below the DTM in the denser portion of the score distribution (raw scores 66 to 80), which partially covered the cut-score range. However, the 95% CI became wider at the ends of the score range, so that the 95% CI covered the DTM. This result indicates that sample sizes in the lower and upper parts of the distribution are insufficient to conclude that the population RMSD values exceed the DTM in these regions, although in this sample they appear to. Figure 10 presents the same information under the equal weight
condition. The pattern of results is almost identical regardless of the weights imposed on the two subgroups.

Figure 11 presents the 95% CI for RMSD with repeater (reference) and repeater (other forms) subgroups differentiated and each subgroup weighted by its proportion in the total group. Although the 95% CI band became narrower for the cut-score range, the band was much wider for raw scores over 80, indicating substantial fluctuation across samples. As shown in Figure 12, the fluctuation was much more pronounced when all subgroups were equally weighted. The DTM values fell outside of the 95% CI band across nearly all the data points. The DTM values were just within the 95% CI for raw scores from 62 to 74. Results indicated subgroup dependence of equating functions when subgroups were defined by the previous form taken.

Discussion

The present study investigated the sensitivity of equating to subgroups based on repeater status. This study examined the conditional RMSD and overall REMSD measures to assess score equatability, along with subgroup specific RESD measures. The overall REMSD was smaller than the DTM, indicating negligible equating differences between the subgroups and the total group when repeaters were not differentiated according to their prior exposure to the reference form. The curvilinear transformation was maintained reasonably well across repeaters and first-timers under this condition.

When we redefined the repeaters with respect to their previous experience with the common anchor items, however, subgroup independence did not hold for the equating function. The equating function derived with repeaters who saw the reference form
previously differed substantially from that derived with the total sample of examinees. The RESD value was very large when the equating function derived using the repeater (reference) subgroup was compared to the total group equating function. The \( e_{wREMSE} \), summarized across the cut-score region, was much larger than DTM mainly due to the repeater (reference) subgroup.

In most score equity assessment analyses, evidence of population dependence might suggest the need to re-evaluate test assembly specifications or linking methods. The results of this study, however, suggest a different remedy. Comparison of repeaters and first-timers in general revealed no differences in the equating functions across subgroups. Thus, familiarity with the material in general or with the test format did not have an impact on the equating function. Only when those individuals with previous exposure to the common items on the test form were examined did differences in equating functions emerge. Regression analyses, coupled with equating results, indicated that the prior exposure to the common items changed the items’ statistical characteristics, thereby invalidating them as anchor items. Here, the indicated problem is neither with the test specifications nor with the linking method, but with a group of individuals with prior knowledge of the test items themselves.

As shown in this study, repeaters’ previous exposure to common items tends to pull the equating function away from the total group equating function. Because the repeaters show enhanced performance on the common items but not on the rest of the test, the new form sample would look more able than the reference form sample, while the new form would look more difficult. This tendency would increase as the proportion of repeaters in the equating sample increased. Such a situation would prove particularly
problematic for licensure tests, because the net effect would be a continued lowering of the raw scores that map to the qualification standard. Clearly such repeaters should be excluded from the equating sample.

As described above, the repeater issue often is discussed in reference to exposure to common items. The assumption that repeaters perform better on all anchor items due to their previous exposure, though, seems unrealistic. Some items might be easier to memorize than others. It might be useful to carry out some statistical checks to discover which anchor items function differently across repeater and first-timer subgroups after adjusting for subgroup members’ differences in ability. Excluding problematic items from the anchor set might be an alternative way to reduce statistical equating error rather than excluding repeaters from the equating sample. If a large proportion of common items function differently across the two subgroups, however, the removal of many common items will bias the final equating function. Excluding repeaters from the equating sample might be the better choice in this situation.

This study has some practical implications, but it also has limitations. First, the sample sizes for the subgroups were highly unbalanced, with the proportion of the repeater (reference) subgroup being particularly small. Consequently, some extremely large RMSD, REMSD, or RESD values emerged partially due to sampling variability. Although the bootstrap procedure was used in this application to compute the standard error of RMSD values to determine whether the differences obtained with the original data were due to random error, implementing this procedure as a standard operation may not be practical for testing programs with strict deadlines. Second, in an operational setting, repeater information and other demographic variables are reported by the
examinees themselves. Examinees may report inaccurate information, and verifying the information they provide is tedious work.

For score equatability, equating must operate independently of subgroups of examinees from whom the data are gathered to develop the conversions. If the equating functions derived using different subgroups differ systematically, the interchangeability of test forms is questionable. Subpopulation invariance might be valid in some situations but not in others because this property is test specific, depending on the definition of subgroups and the characteristics of the ability being measured. Invariance of a certain subgroup on a certain test cannot be generalized to either other subgroups or other tests. This means that invariance investigations should be conducted periodically with major subgroups for each test to ensure the fairness of score reporting, particularly for large-volume, high-stakes tests.
References


### Table 1

**Summary Statistics for the New and Reference Form Examinees Groups**

<table>
<thead>
<tr>
<th>Test Form</th>
<th>Group</th>
<th>N</th>
<th>(%)</th>
<th>Total</th>
<th>Anchor</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>r</td>
</tr>
<tr>
<td>New</td>
<td>Total</td>
<td>6426</td>
<td>(100%)</td>
<td>73.62</td>
<td>10.51</td>
<td>30.60</td>
<td>4.96</td>
<td>.91</td>
</tr>
<tr>
<td>Form X</td>
<td>First-timer</td>
<td>4211</td>
<td>(66%)</td>
<td>77.12</td>
<td>9.55</td>
<td>32.16</td>
<td>4.45</td>
<td>.89</td>
</tr>
<tr>
<td></td>
<td>Repeater</td>
<td>2215</td>
<td>(34%)</td>
<td>66.95</td>
<td>8.88</td>
<td>27.62</td>
<td>4.51</td>
<td>.86</td>
</tr>
<tr>
<td></td>
<td>Reference</td>
<td>330</td>
<td>(5%)</td>
<td>64.66</td>
<td>8.54</td>
<td>26.81</td>
<td>4.57</td>
<td>.86</td>
</tr>
<tr>
<td></td>
<td>Other forms</td>
<td>1885</td>
<td>(29%)</td>
<td>67.35</td>
<td>8.88</td>
<td>27.76</td>
<td>4.48</td>
<td>.86</td>
</tr>
<tr>
<td>Reference</td>
<td>Total</td>
<td>6489</td>
<td>(100%)</td>
<td>77.47</td>
<td>10.83</td>
<td>30.46</td>
<td>5.09</td>
<td>.92</td>
</tr>
<tr>
<td>Form Y</td>
<td>First-timer</td>
<td>4117</td>
<td>(63%)</td>
<td>81.90</td>
<td>9.19</td>
<td>32.49</td>
<td>4.29</td>
<td>.90</td>
</tr>
<tr>
<td></td>
<td>Repeater</td>
<td>2372</td>
<td>(37%)</td>
<td>69.79</td>
<td>9.00</td>
<td>26.93</td>
<td>4.40</td>
<td>.86</td>
</tr>
<tr>
<td></td>
<td>Reference</td>
<td>470</td>
<td>(7%)</td>
<td>69.26</td>
<td>8.62</td>
<td>26.20</td>
<td>4.21</td>
<td>.85</td>
</tr>
<tr>
<td></td>
<td>Other forms</td>
<td>1902</td>
<td>(29%)</td>
<td>69.92</td>
<td>9.09</td>
<td>27.12</td>
<td>4.43</td>
<td>.87</td>
</tr>
</tbody>
</table>

*Note.* Repeater (Reference) indicates repeaters who took the exact reference form at the previous administrations. Repeater (Other forms) indicates repeaters who took any form other than the reference form at the previous administrations. $r$ indicates the correlation between the total and anchor scores.
Figure 1. Relative frequency distribution of new form X scores in the three subgroups.
Figure 2. The general framework for equating new form X to reference form Y
Figure 3. Separate regression lines for repeater (reference) and repeater (other forms) in the new Form X group
Figure 4. Separate regression lines for repeater (reference) and repeater (other forms) in the reference Form Y group
Figure 5. Equating functions derived using total and various subgroup equating samples
Figure 6. Difference curves between the total group equating function and subgroup equating function.
Figure 7. Score-level RMSD and overall REMSD derived from the comparison of the total group equating function to the two subgroup equating functions.
Figure 8. Score-level RMSD and overall REMSD derived from the comparison of the total group equating function to the three subgroup equating functions.
Figure 9. The 95% confidence interval band of RMSD with two subgroups weighted unequally based on their proportion in the total group.
Figure 10. The 95% confidence interval band of RMSD with two subgroups weighted equally
Figure 11. The 95% confidence interval band of RMSD with three subgroups weighted unequally based on their proportion in the total group
Figure 12. The 95% confidence interval band of RMSD with three subgroups weighted equally.