Research Report

Single- Versus Double-Scoring of Trend Responses in Trend Score Equating With Constructed-Response Tests

Xuan Tan

Kathryn L. Ricker

Gautam Puhan

April 2010

ETS RR-10-12

Single- Versus Double-Scoring of Trend Responses in Trend Score Equating With Constructed-Response Tests

Xuan Tan, Kathryn L. Ricker, and Gautam Puhan
ETS, Princeton, New Jersey

April 2010

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: Longjuan Liang and Jinghua Liu

Abstract

This study examines the differences in equating outcomes between two trend score equating designs resulting from two different scoring strategies for trend scoring when operational constructed-response (CR) items are double-scored—the single group (SG) design, where each trend CR item is double-scored, and the nonequivalent groups with anchor test (NEAT) design, where each trend CR item is single-scored during trend score equating—for varying sample sizes (n =150, 200, 250, 300, 400). Overall results suggest larger equating errors with smaller sample sizes, though errors were small regardless of sample size. The NEAT design performed about as well as the SG design with respect to conditional and summative standard errors of equating, though it did tend to produce larger bias and root mean-squared differences (RMSDs). When accounting for the total number of trend scores required to do analyses, the NEAT design performed as well or better than the SG design (e.g., when the NEAT n =150 and the SG n = 300). This result might be partially attributable to a larger operational sample size (n = 792) and a good correlation between anchor and total score for the trend sample (r = 0.73). These results suggest that under these testing conditions, the NEAT design performed about as well as the SG design, but further research is required to assess the generalizability of the results.

Key words: trend scoring, trend-score equating, constructed response, quality

Acknowledgments

Many thanks to Matthew Duchnowski, Leanne Gall, Danielle Siwek, and Brian Sucevic for their invaluable work in running the original analyses. Thanks to Sooyeon Kim for her advice and programs used to run the bootstrap sample conditions. Thanks to Tim Moses for his extensive programming assistance. Thanks to Michael Walker for his advice and input throughout the project. Thanks to Dan Eignor for his edits and feedback on an earlier version of this manuscript.

When a multiple-choice (MC) test form is reused (i.e., as a reprint form), the original raw-to-scale conversion is often applied to the reprint form. However, when a constructed-response (CR) test form is reused, applying the original raw-to-scale conversion to the reprint form may not be appropriate if the effective scoring standards from the original and reprint administrations are different. This discrepancy occurs because human raters (i.e., scorers), despite their best efforts, are often not successful in applying identical scoring standards over two time points. Therefore it is necessary to evaluate whether the scoring standards from the original and reprint administrations are the same. If they are not the same, then an adjustment to the original raw-to-scale conversion is necessary before it can be applied to the reprint form. How can this adjustment be made?

The answer lies in *trend scoring and equating* (Livingston, 2007), a method wherein some or all of the examinee responses from the original CR form are rescored during the reprint administration along with the examinee responses from the reprint form. This rescoring leads to two sets of ratings for the original responses: ratings assigned by raters in the original administration and ratings assigned by the raters in the reprint administration. The scores based on the rescoring are known as trend scores. These two sets of ratings can be compared to evaluate if the scoring standards have changed between the original and reprint scoring time points. If the scoring standards are similar, then the original raw-to-scale conversion can be applied to the reprint form. If the scoring standards are different, then equating is needed to create a new raw-to-scale conversion table for the reprint form (see Kamata & Tate, 2005; Tate, 1999, 2000, 2003, who used a similar procedure for long-term scale maintenance in an item response theory [IRT] context).

A factor that needs to be considered when using the trend scoring method is whether to single or double score (i.e., rate once or twice) the original responses during the trend scoring. If the original responses were single-scored, then using the same process for the rescoring (i.e., single scoring) seems reasonable. But if the original responses were double-scored, then should the trend sample be single- or double-scored? Although using two sets of raters in the rescoring when the original responses were double-scored is ideal (because it better matches the original scoring process), it is also more expensive and time-consuming than single scoring. On the other hand, single scoring the original responses along with double-scored responses from the reprint form may be cumbersome logistically. For example, the interspersing[1] of the regular examinee

responses with the examinee responses from the reprint administration becomes more complex if one set of responses have to be single-scored and another set of responses have to be double-scored. This study will therefore examine whether single or double scoring of the trend responses leads to similar equating results (i.e., lead in both cases to the same decision that the original raw-to-scale conversion table can or cannot be used in the reprint form).

Single versus double scoring of the trend papers has important implications for the equating designs that can be used to adjust for differences in scoring standards between the original and reprint forms. Using the same number of ratings in the trend scoring as in the original scoring facilitates the use of a single group (SG) equating design, while using a different number of ratings in the trend scoring as compared to the original scoring makes it possible to use the nonequivalent groups with anchor test (NEAT) design but not the SG design. These two equating scenarios are described next.

Single Group (SG) and Nonequivalent Groups With Anchor Test (NEAT) Equating Scenarios

For the purpose of illustration, consider Form X as the original form and Form X-R as the reprint form. If the scores on Form X and the scores on the trend sample (i.e., a sample of Form X responses that are rescored in the Form X-R administration) are rated by the same number of raters (e.g., double-scored for both), then the trend scores can be equated to the original scores using an SG equating design to adjust for changes in scoring standards. Since the responses came from the same examines with the same number of ratings and the only difference is that the same responses are scored by raters from two time points, an SG equating design can be used. The resulting conversion can then be applied to the reprint form since the trend scores and the reprint form (Form X-R) responses are based on the same items and rated by the same set of raters (see Figure 1 for illustration).

However, if the scores on Form X and scores on the trend sample are not rated by the same number of raters (e.g., original scoring used double ratings while the trend scoring used a single rating), then trend scores cannot be equated to the original scores using an SG equating design. Instead a NEAT equating design has to be used. Under this design, Form X (treated as the reference form) total scores will be composed of double-scored responses of examinees who took Form X, and Form X anchor scores will be comprised of single-scored responses of

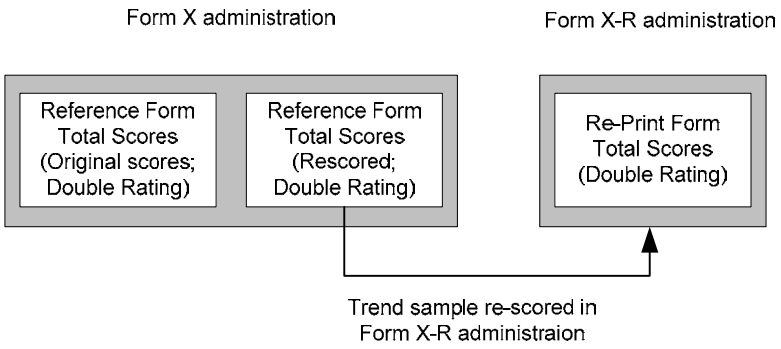


Figure 1. **Illustration of the single group (SG) equating design where trend scores and operational scores are both based on double ratings.**

Note. Form X = original form; Form X-R = reprint form.

examinees who took Form X but whose responses were rescored during the Form X-R administration. Therefore, although the total and anchor scores were based on the same items, they were obtained from two different sets of raters (one set from the Form X administration and one set from the rescoring session in Form X-R administration). This makes the anchor scores somewhat external to the total scores. Form X-R (treated as the new form) total scores will be comprised of double-scored responses of examinees who took Form X-R, and Form X-R anchor scores will be comprised of the first rating of all the examinee responses who took Form X-R. The actual number of items included in the total and anchor scores is, again, exactly the same. The only difference is in the number of ratings where the anchor scores include a single rating but the total scores include double ratings. Even though the anchor score for Form X-R actually has two ratings, only one rating (either the first or second rating) is used to match the single rating anchor scores on Form X (see Figure 2 for illustration). The anchor scores, in this case, were obtained from the same set of raters in the Form X-R administration and contributed to the total scores. Thus, the anchor scores are internal to the total scores.

Purpose of the Study

The study examined the impact of single versus double scoring of trend responses on final equating results of a reprint form. The data examined in this study is from a testing program where all operational scores are based on double ratings. The testing program wanted to make a decision whether to single or double score the trend papers and to find out the minimum sample

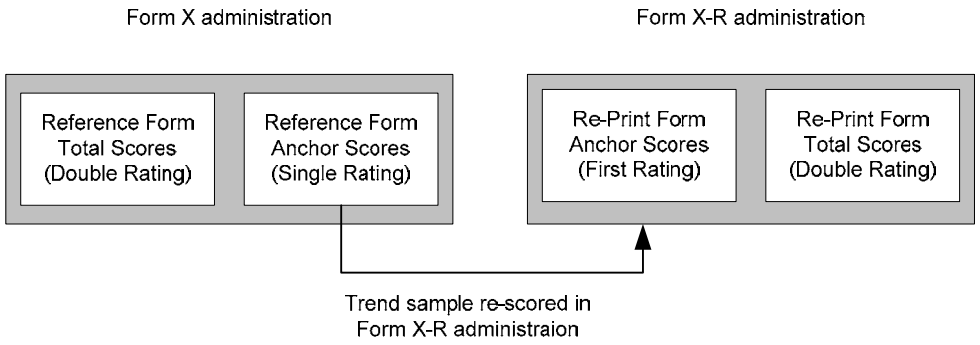


Figure 2. **Illustration of the nonequivalent groups with anchor test (NEAT) equating design where trend scores are based on single ratings but operational scores are based on double ratings.**

Note. Form X = original form; Form X-R = reprint form.

size required for the trend scoring in order to conduct an acceptable equating under both conditions. Specifically, the purpose of the study is to examine the following questions:

1. How many single-scored trend responses are needed to obtain an acceptable equating for the reprint form?
2. How many double-scored trend responses are needed to obtain an acceptable equating for the reprint form?
3. Since double scoring the trend responses leads to logistical advantages (see the introduction) and theoretical advantages (SG design has been shown to have much less error than the NEAT design; see Thorndike, 1982, and Kolen & Brennan, 2004), can fewer double-scored trend responses as compared to more single-scored trend responses be used to conduct the equating (e.g., 200 double-scored versus 400 single-scored trend responses)?

Method

Data Collection and Analysis

The study used test data from a large-scale certification test that consisted of four CR items resulting in 48 score points (4 items × 6 maximum points per item × 2 ratings). Throughout the paper, the original or previously used form of the test will be referred to as Form

X and the reprint form will be referred to as Form X-R. Form X was first administered in 2005, and Form X-R was administered in 2006. A total of 452 examinee responses from Form X (originally scored in 2005) were interspersed with 792 responses of examinees who took Form X-R in 2006, and all responses were double-scored using the 2006 raters. Although the sample of examinees that took Form X in 2005 was larger than 452, only 452 examinee responses could be trend-scored (i.e., rescored) because of financial and logistical constraints. Since these 452 examinee responses were double-scored during the trend scoring and the original scoring also used double scoring, the trend scores could be equated to the original scores using an SG equating design. The resulting conversion was then applied to Form X-R. This conversion will be considered the *criterion equating* to which equatings based on smaller samples of trend-scored responses (either single- or double-scored) will be compared to evaluate whether a smaller number of trend-scored papers can be used to conduct an acceptable equating. Although all the trend responses were double-scored, only the first or third rating (which existed if adjudication of the first two ratings was needed) was used to mimic and evaluate the single trend-scored condition (i.e., the NEAT design with single scores).

For the NEAT design, Form X-R was considered as the new form and Form X was considered as the reference form. Form X-R total scores were comprised of 792 responses (double-scored) from the Form X-R administration, and Form X-R anchor scores were comprised of 792 responses (first or third rating only) from the Form X-R administration. Form X total scores were comprised of 452 responses (double-scored) from the Form X administration, and Form X anchor scores were composed of 452 responses (first or third rating only) rescored during the Form X-R administration. The sample sizes examined were 150, 200, 250, and 300. These sample sizes were varied only for the Form X responses (i.e., the reference sample). Form X-R sample sizes were unaltered because in a real testing situation all the operational responses of examinees who take the new form (i.e., Form X-R) have to be scored. Hence, the question of cost savings by making the Form X-R sample smaller does not arise.[2] For a particular sample size condition (e.g., 150), a random sampling with replacement procedure was used to select 150 examinee responses from the 452 responses. Then a NEAT equating was conducted using these 150 examinee responses as the reference form sample and 792 examinee responses as the new form sample. This process was repeated 200 times for each sample size condition to calculate estimates of equating error and bias. The equating methods used in the NEAT condition were

chained linear and chained equipercentile equatings.[3] When comparing the individual equatings with the criterion, linear equatings were compared with a criterion linear equating and nonlinear equatings were compared with a criterion nonlinear equating.

For the SG design, the adjustments for emulating the single-scored NEAT condition were not needed. The SG equatings involved equating the Form X original responses (scored using two sets of raters) directly to Form X rescored responses (scored using two set of raters as well). Sample sizes of 150, 200, 250, and 300 were also examined for the SG equatings. Similar to the NEAT condition, a random sampling with replacement procedure was used to select a particular number of responses (e.g., 150) from the 452 responses. Then a SG equating was conducted using the rescored responses as the new form sample and the original responses as the old form sample. This process was also repeated 200 times for each sample size condition to calculate estimates of equating error and bias. The equating methods used in this condition were direct linear and direct equipercentile equatings.[3]

Logistically, double scoring means that twice the number of trend papers needs to be rescored compared to single scoring for the same sample size. Thus, a fair comparison would be to evaluate the effectiveness of double scoring compared to single scoring with half the number of trend papers used in single scoring. With the different sample size conditions in the study, one such comparison can be made between the SG design with a 150 double-scored trend sample and the NEAT design with a 300 single-scored trend sample. To make the comparison more generalizable, one more sample size condition of 400 was simulated for the single-scored NEAT design condition to make possible another comparison between the SG design with a 200 double-scored trend sample and the NEAT design with a 400 single-scored trend sample.

Variability and Accuracy Indexes

For both the single-scored NEAT and double-scored SG conditions, the standard deviation of the 200 equatings provided an estimate of random equating error. This error was calculated at individual score levels and was referred to as the *conditional standard error of equating* (CSEE). The CSEEs were then summed up and averaged to get a summary statistic at the total test score level and were then referred to as the *average standard error of equating* (SEE). The average difference between the individual equatings and the criterion equating across the 200 replications provided an estimate of equating bias. This bias was also calculated at the individual score levels and referred to as *conditional bias*. When calculated at the total test score

level, it was referred to as *bias*. Finally, at the total test score level, both the error and bias estimates were used to provide an index of overall equating accuracy, known as the root mean-squared difference (RMSD; see Puhan, Moses, Grant, & McHale, 2008, for details.)

For all calculations, the total score range was truncated to 24 to 40 points. This truncation was done because 90% of the trend sample fell into that score range. The sparseness of data in the other parts of the score range would likely cause error estimates to be highly inflated.

In order to interpret the results, an interpretive criterion was needed. The practical criterion of the difference that matters (DTM; Dorans & Feigenbaum, 1994) was used. Briefly, the DTM is a unit equal to one-half of a reporting unit. In this case, where raw scores are used, it is 0.5 of a raw score unit. Any error larger than that would be noticeable, that is, would matter.

Results

Results from the study are presented in the following sections. First, statistical characteristics of the equating samples are summarized across different conditions. Then, the different indexes for evaluating equating efficiency are presented. Results for the CSEE and overall SEE are presented first, followed by results for the conditional bias and the overall bias. Results for the average RMSD are presented last. These results are compared across different studied conditions to shed light on how equating design (NEAT or SG, due to different scoring for the trend sample—single and double scoring) and sample size affect the accuracy of trend score equating. Finally, average SEE, bias, and RMSD results are compared holding the number of trend papers being rescored equal for the two designs (150 double-scored versus 300 single-scored, and 200 double-scored versus 400 single-scored trend papers).

Statistical Characteristics of Equating Samples

For the double-scored SG design, the means and standard deviations (SD) of the total scores based on the original scoring and rescoring of the trend samples were calculated for the total sample (452 trend sample) and the smaller samples across all simulated conditions. For the different sample size conditions, the means and SDs of total scores were averaged across the 200 replications to provide a summary for each condition. These statistics turned out to be fairly close to those calculated for the total sample. The differences were all within 0.05. For the single-scored NEAT design, the means and SDs of the total and anchor scores, as well as the correlation between the total and anchor scores, were calculated for the total sample (792 new

form, 452 trend samples) and the simulated samples. The averages across 200 replications in each sample size condition were again calculated and compared to those obtained from the total sample. The statistics again turned out to be fairly close with differences smaller than 0.02. Tables 1 and 2 present the statistics calculated for the total samples for the two designs.[4]

Table 1

Statistical Characteristics of the Equating Sample (Single Group [SG] Design—452 Trend Sample)

	Rescored total	Original total
Mean	33.44	33.53
SD	3.65	3.22

Table 2

Statistical Characteristics of the Equating Samples (Nonequivalent Groups With Anchor Test [NEAT] Design—792 New and 452 Reference Samples)

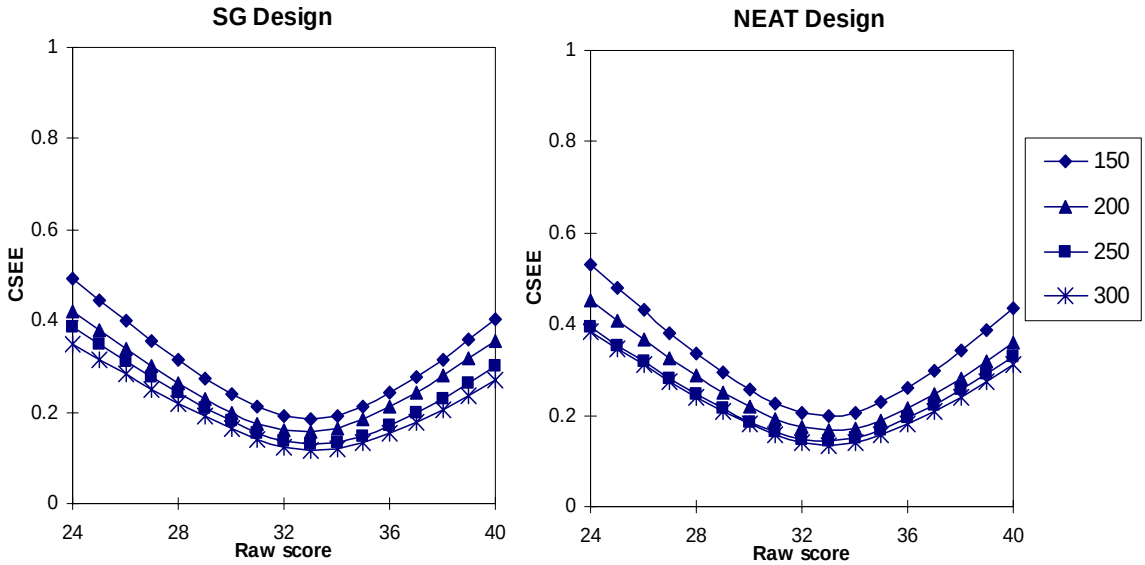
	New form	Reference form
Total score		
Mean	32.50	33.53
SD	3.87	3.22
Anchor score		
Mean	16.25	16.70
SD	2.00	1.95
Correlation between total/anchor[a]	0.96	0.73

[a] The correlation between total and anchor scores is usually much higher for the new form than for the reference form. This difference occurs because for the reference form the total and the anchor scores are rated by different sets of raters and thus have a weaker link between them (the anchor is the single rating obtained in the rescoring of trend papers in the reprint form administration).

Conditional Standard Errors of Equating (CSEE)

Figure 3 contains results of the CSEEs for the linear equating methods. The CSEEs obtained from the direct linear equating under the SG design condition are plotted in Figure 3(a). The CSEEs obtained from the chained linear equating under the NEAT design condition are plotted in Figure 3(b). Figure 4 contains results of the CSEEs for the nonlinear equating methods. Similarly, Figure 4(a) presents the results for the direct equipercentile equating in the SG design, and Figure 4(b) presents the results for the chained equipercentile equating in the NEAT design.

Overall, CSEEs for the linear equating methods were small across the score scale (24 to 40). The CSEEs were smaller than a DTM for all except one condition (NEAT design with a sample size of 150). The NEAT and SG designs tended to perform very similarly, and the errors for both designs tended to be incrementally smaller as the sample size increased from 150 to 300. As expected, the errors are smallest overall in the region where the most data is located (i.e., around the mean of the test scores or approximately 33-34).

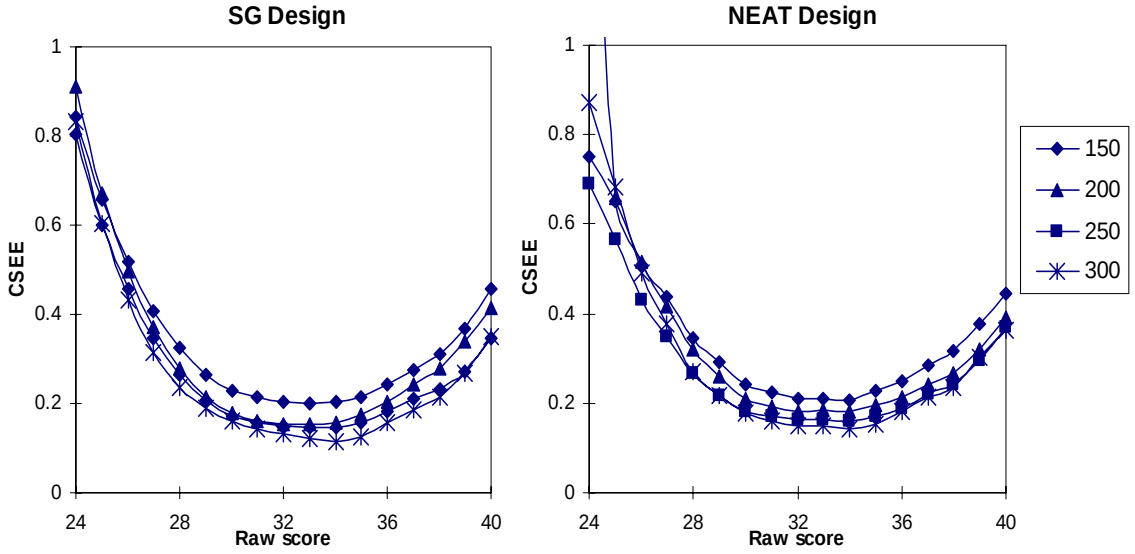


3(a) 3(b)

Figure 3. **Conditional standard errors of equating (CSEEs) across equating designs and sample sizes for linear equating methods.**

Note. SG = single group; NEAT = nonequivalent groups with anchor test.

As shown in Figure 4, CSEEs for nonlinear methods of equating are larger than a DTM at score points below 26 but smaller than a DTM for all other score points. The same pattern of results for both the SG and NEAT designs can be observed, with the exception of the very lowest scores (24 to 26) where the NEAT design tends to have smaller errors for the 250 condition. Moreover, the errors for those score points are smaller for the 150 and 250 SG conditions than they are for the 300 conditions. Aside from this aberration at score points from 24 to 26, the same pattern of smaller errors with larger sample sizes was observed in the nonlinear equating methods as was observed in the linear equatings.



4(a) 4(b)

Figure 4. **Conditional standard errors of equating (CSEEs) across equating designs and sample sizes for nonlinear equating methods.**

Note. SG = single group; NEAT = nonequivalent groups with anchor test.

Average Standard Errors of Equating (SEE)

Figure 5 presents the results of the average SEEs for the direct linear and direct equipercentile equatings in the SG design and for the chained linear and chained equipercentile equatings in the NEAT design across sample sizes. The average SEEs for all designs and methods were smaller than a DTM across all sample sizes. The SG design provided smaller average SEEs than the NEAT design across all sample size and equating method conditions.

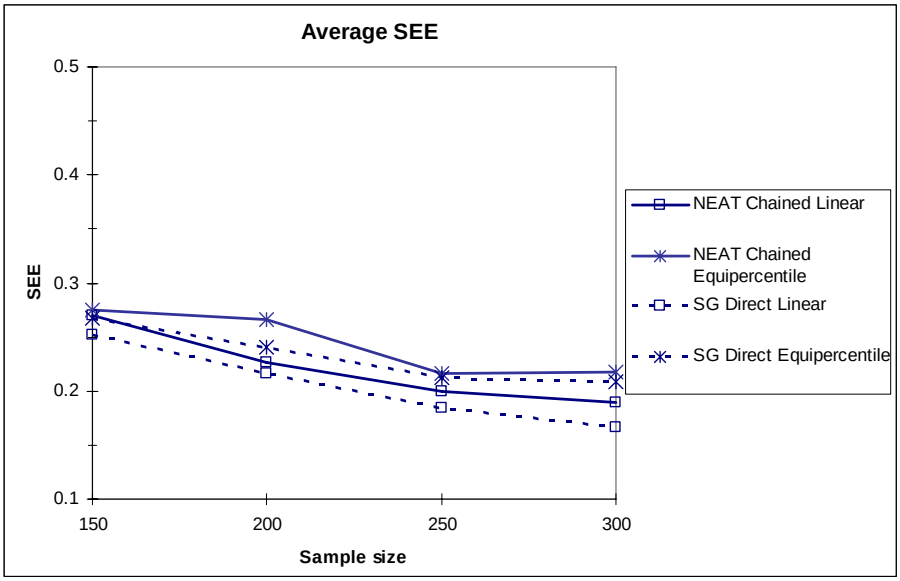


Figure 5. **Average standard equating errors (SEEs) across equating designs, equating methods, and sample sizes.**

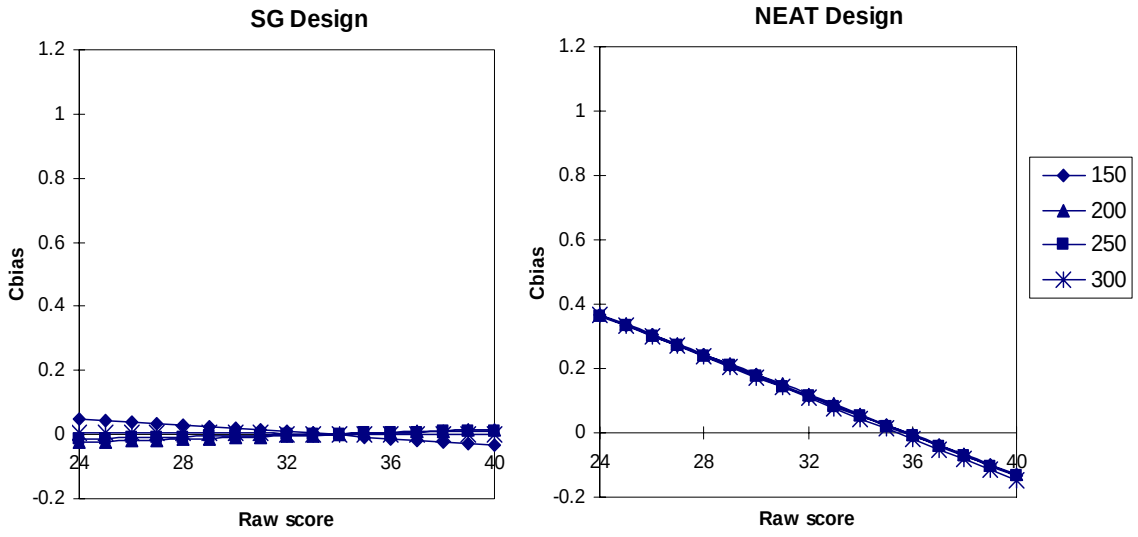
Note. SG = single group; NEAT = nonequivalent groups with anchor test.

The linear methods performed similarly to each other, as did the nonlinear methods. Overall, the linear methods tended to have smaller average SEEs. Average SEEs for all equating designs and methods tended to decrease as the sample size increased.

Conditional Bias

Figures 6 and 7 contain results of the conditional bias across studied conditions for the linear and nonlinear equating methods respectively. Each figure contains two graphs, (a) and (b), which present results for the SG design and NEAT design, respectively.

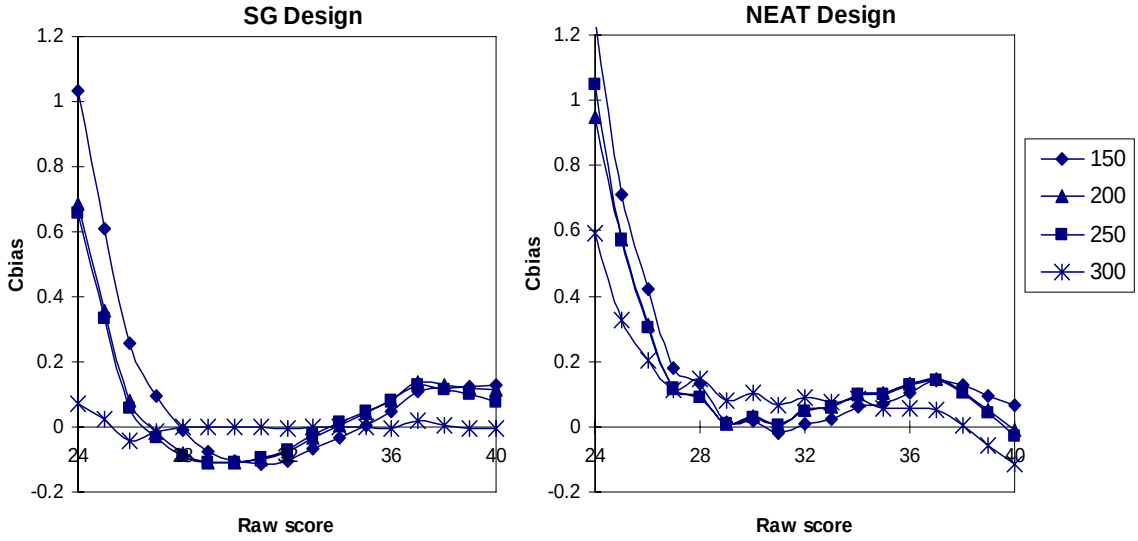
Conditional biases for the linear equating methods were very small across the score scale (Figure 6). They were all within the range of -0.2 to 0.4 (smaller than a DTM). Sample size showed a slight influence on conditional bias for the SG design with smaller conditional biases obtained for larger sample sizes. Sample sizes did not show any significant influence on conditional biases for the NEAT design since the different lines for different sample sizes were basically on top of each other. The SG design had smaller conditional biases than the NEAT design across the whole score scale, an observation that can be partially attributed to the fact that the criterion equating function was based on the SG equating.



6(a) 6(b)

Figure 6. **Conditional biases (CBias) across equating designs and sample sizes for linear equating methods.**

Note. SG = single group; NEAT = nonequivalent groups with anchor test.



7(a) 7(b)

Figure 7. **Conditional biases (CBias) across equating design and sample size for nonlinear equating methods.**

Note. SG = single group; NEAT = nonequivalent groups with anchor test.

Conditional biases for the nonlinear equating methods were fairly large, larger than a DTM for the very low end of the score scale (24 to 26) where data was sparse (see Figure 7). Beyond the score point of 26, conditional biases were all within the range of -0.2 to 0.5. Sample size showed some slight influence at the low end of the score scale (24 to 26), where as sample size increased, conditional bias decreased. However, for the rest of the score scale, sample size did not show any impact on conditional bias since the lines were either on top of each other or crossing at various points. Overall, the SG design had smaller conditional biases than the NEAT design across the whole score scale.

Bias

Figure 8 displays the overall equating bias for the direct linear and direct equipercentile equatings in the SG design and for the chained linear and chained equipercentile equatings in the NEAT design across sample sizes. As shown in the graph, the summarized biases were much smaller than a DTM across all conditions. The lines were basically horizontal, indicating no influence of sample size on bias. Within each equating design, the linear and nonlinear conversions were almost on top of each other. The biases were very similar across different equating methods within each equating design. The SG design consistently had smaller biases across sample size and equating method conditions, although the differences were relatively small, around 0.1.

Root Mean-Squared Difference (RMSD)

Figure 9 displays the RMSDs, which provide summary information on the combination of random SEEs and systematic equating bias. Similar trends were identified across equating methods. The RMSDs were very small across all conditions (< 0.1). A slight declining trend showed across sample size conditions as sample size increased. The SG design outperformed the NEAT design producing smaller RMSDs within each equating method condition. The line providing the smallest RMSDs is the condition with SG design and the direct linear equating method. The differences in RMSD were again too small to indicate a significant impact of different equating designs on the accuracy of the equating.

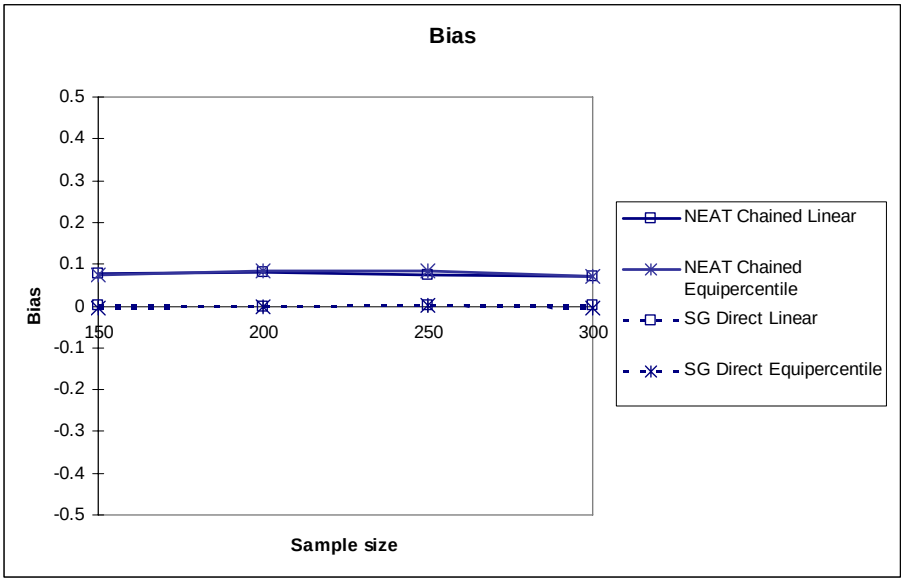


Figure 8. **Biases across equating designs, equating methods, and sample sizes.**

Note. SG = single group; NEAT = nonequivalent groups with anchor test.

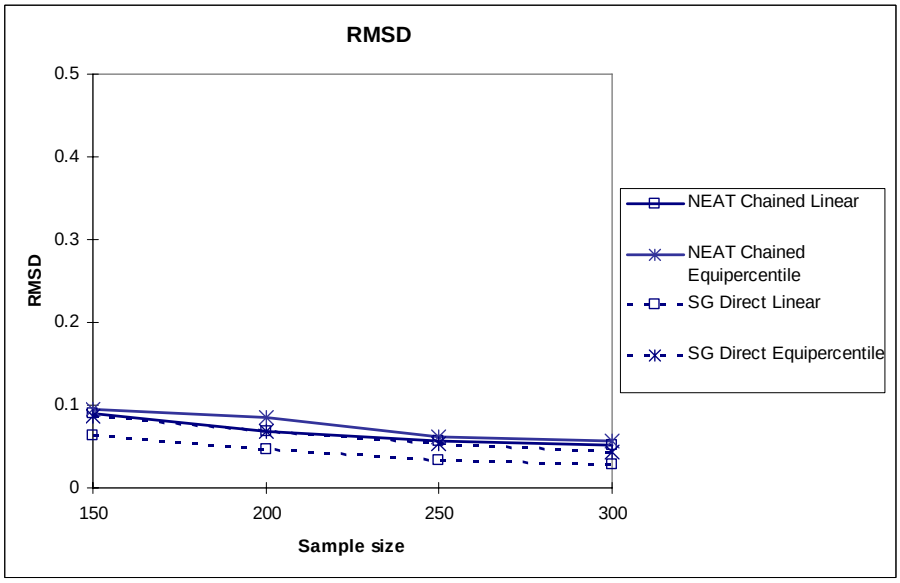


Figure 9. **The root mean-squared differences (RMSDs) across equating designs, equating methods, and sample sizes.**

Note. SG = single group; NEAT = nonequivalent groups with anchor test.

Efficacy of the Double-Scored Single Group (SG) Design and the Single-Scored Nonequivalent Groups with Anchor Test (NEAT) Design

Figures 10 and 11 contain the comparison results for the efficacy of the two equating designs (due to different scoring schema for the trend samples). The efficacy was evaluated in two ways: (a) the amount of rescoring needed for the trend sample, and (b) the equating accuracy. Two comparisons were made: (a) the SG design with the 150 double-scored trend sample versus the NEAT design with the 300 single-scored trend sample (Figure 10), and (b) the SG design with the 200 double-scored trend sample versus the NEAT design with the 400 single-scored trend sample (Figure 11). The average SEE, bias, and RMSD were compared across the two designs.

Figure 10 shows that, with the same amount of rescoring work (300 rescores required for 150 trend papers with double ratings or for 300 trend papers with single ratings), the NEAT design obtained slightly smaller average SEEs and RMSDs for both linear and nonlinear equating methods. However, the SG design obtained slightly smaller biases for both linear and nonlinear equating methods. Since all values were smaller than a DTM and the differences were all within 0.1, both designs worked equally well.

Discussion

Effect of Sample Size

Overall, an effect of trend sample size was observed. For all variables we examined in this study, there is a trend towards larger errors as the sample size used for trend scoring was decreased. Within the range of sample sizes that we examined, there was no clear threshold sample size where equating errors became too large. Rather, within the range of scores reported in the results section (which includes the passing cut score range for this test title), the errors tended to be very small, regardless of sample size.

However, sample size did have an interesting effect on the equating decisions we made across replications. In our simulation, the decision to equate was based on the examination of the total trend sample of 452 papers. If we were to reevaluate when each sample was drawn whether equating was warranted, then we might have had replications where we would have decided that equating was not needed. A reexamination was done for all 200 replications within each sample size condition by examining whether the differences between the linear equating lines using the SG design and the identity lines were larger than two SDs of equating errors for the score range

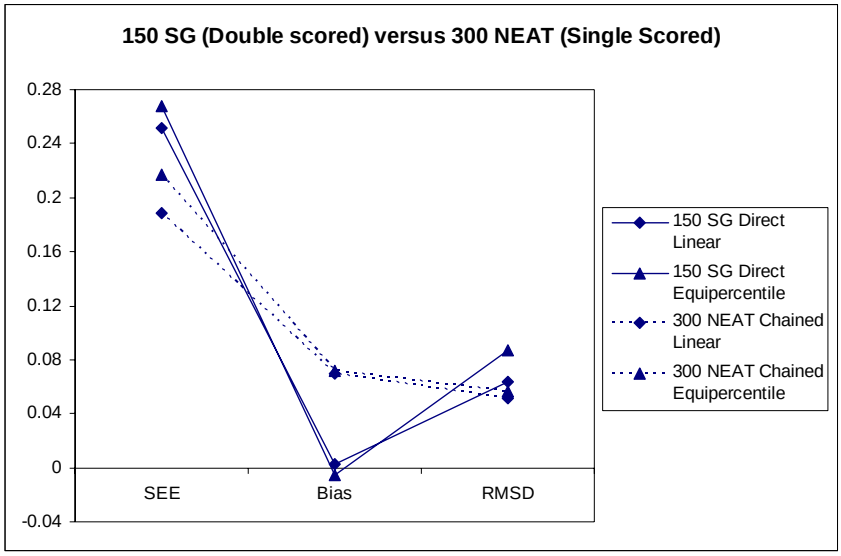


Figure 10. **Average standard errors of equating (SEEs), biases, and root mean-squared differences (RMSDs) for the 150 single group (SG) design and 300 nonequivalent groups with anchor test (NEAT) design.**

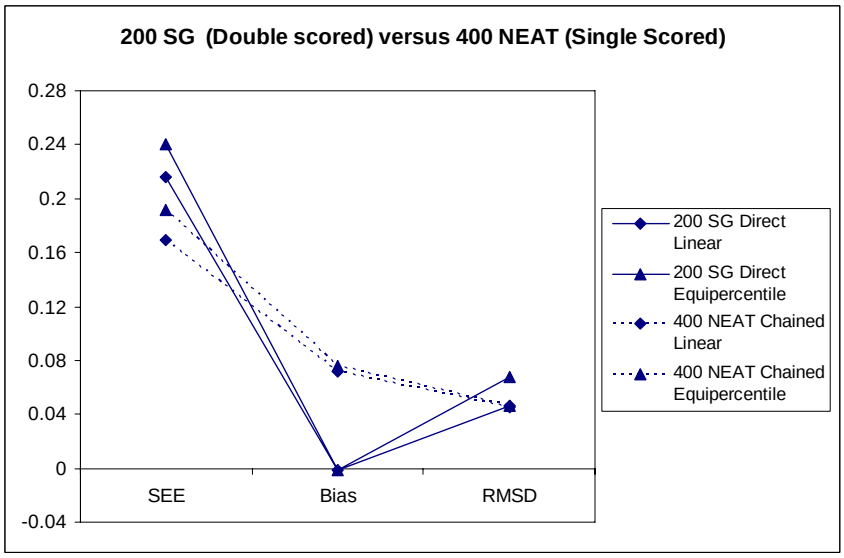


Figure 11. **Average standard errors of equating (SEEs), biases, and root mean-squared differences (RMSDs) for the 200 single group (SG) design and 400 nonequivalent group with anchor test (NEAT) design.**

of 24 to 40. The percentage of replications where the same equating decision (yes to equate) was made for the 150, 200, 250, and 300 sample size conditions were 80%, 87.5%, 88.5%, and 94.5% respectively. The 300 sample size condition seemed to have the most appropriate agreement rate with the equating decision made with the total trend sample (close to 95%). This result, however, should be interpreted with caution. First, the criterion of identifying differences larger than two SDs of error across the score scale is fairly stringent. Given that data was sparse at the extreme ends of the score scale, equating errors were fairly large, making it more difficult to exceed two SDs of equating error. Second, this criterion may not agree with other criterions used in practice (e.g., DTM).

Effect of Single Group (SG) Versus Nonequivalent Groups With Anchor Test (NEAT) Design

The NEAT design performed as well as the SG design with regards to both conditional and summative SEEs. The SG design did produce less bias and smaller RMSDs. However, the better performance of the SG design could also be at least somewhat attributable to the fact that our selected criterion was the SG design, not the NEAT design. The efficacy comparisons (done by holding the total number of rescores equal) reached similar conclusions. The NEAT design obtained slightly smaller average SEEs and RMSDs and slightly larger biases. However, the differences were too small to conclude which method was superior.

The NEAT design performed better than expected, especially at the smaller sample sizes. We hypothesize that this performance perhaps is due to the combined effect of the NEAT design having the advantage of a larger new form sample (n = 792), along with a relatively high rater1–rater2 correlation in these data (r = 0.78). As the correlation between rater1 and rater2 is increased, less difference would be expected between the results from the SG and NEAT designs, because in this case, two independent scores would not provide much more accuracy to the scoring than only one independent score. Moreover, as shown in the total sample statistics (see Table 2), the correlation between the anchor and the total was relatively high, 0.73, indicating a relatively consistent rescoring with the original scoring.

Comparison of Equating Methods

Overall, the linear methods tended to produce similar results and the nonlinear methods tended to produce similar results. None of these methods produced significantly less or more error or bias than any other method.

Study Limitations

Sparse Data

In order to present more meaningful results, we limited the score scale for the conditional SEE and conditional bias to a range from 24 to 40 (5^{th} and 95^{th} percentile) where there was data. However, the conditional SEE and bias results were still influenced by the fact that only sparse data existed for the score range from 24 to 26. For this score range, the smoothed counts were smaller than 6 for the 452 trend sample, and the smoothed proportions in the total sample were smaller than or equal to 0.01. This contributed to the large numbers for the conditional SEEs and the conditional biases, and to the unexpected pattern of large conditional SEEs for the SG design obtained at the very low end of the score scale. However when the average SEEs and bias were calculated at the total score level, this influence was removed by the averaging effect of the summary statistics.

Inter-Rater Correlation

Another limitation of this study that has already been noted is the interpretation of the effectiveness of the NEAT design versus the SG design. In this sample, the rater1–rater2 correlation is higher than might be expected with a 7-point raw item scale (0–6). Therefore the NEAT design might perform better than would be expected when compared to the SG design. The added benefit of the scoring accuracy provided by two independent raters is lost as the correlation between rater1 and rater2 approaches 1.0. Caution should be exercised in interpreting the performance of the NEAT design in this study.

Final Recommendations

Based on these results, we cannot make any conclusive operational recommendations. Evidence suggests that the SG design holds an advantage over the NEAT design for reducing bias, but that advantage is not to the extent that might have been expected. For SEE, the NEAT design performed about as well as and sometimes better than the SG design, but this performance could be attributed to factors that are specific to the data used in this study. As for the sample size, under this specific set of conditions, smaller sample sizes seemed to perform reasonably within the range of scores reported in this study, but the sparse data at the ends of the score range will be increasingly problematic with smaller sample sizes, so using smaller trend sample sizes operationally could still be a problem. Further research under a variety of conditions is warranted.

References

Dorans, N. J., & Feigenbaum, M. D. (1994). *Equating issues engendered by changes to the SAT® and PSAT/NMSQT®* (ETS Research Memorandum No. RM-94-10). Princeton, NJ: ETS.

Kamata, A., & Tate, R. (2005). The performance of a method for the long-term equating of mixed-format assessment. *Journal of Educational Measurement, 42,* 193–213.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Livingston, S. A. (2007). *Equating and scaling for statistical analysis professionals* [PowerPoint presentation]. ETS: Princeton, NJ.

Puhan, G., Moses, T., Grant, M., & McHale, F. (2008). *An alternative data collection design for equating with very small samples* (ETS Research Rep. No. RR-08-11). Princeton, NJ: ETS.

Tate, R. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal of Educational Measurement, 36,* 336–346.

Tate, R. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement, 37,* 329–346.

Tate, R. L. (2003). Equating for long-term scale maintenance of mixed format tests containing multiple choice and constructed response items. *Educational and Psychological Measurement, 63,* 893–914.

Thorndike, R. L. (1982). *Applied psychometrics*. Boston, MA: Houghton.

Notes

[1] The original responses are interspersed with the responses from the reprint form so that any systematic changes in the way raters score the responses (e.g., change due to fatigue, change due to some raters dropping out on a second day of scoring) affect the scoring of the original and the reprint responses in the same manner.

[2] Although the Form X-R sample size was unaltered, each equating was conducted using a group of 792 examinee responses in the X-R samples selected using a random sampling with replacement procedure, meaning that the 792 responses used in each equating would be slightly different from one another.

[3] Kernel linear and nonlinear equating methods were also examined, but results are not presented because they were very similar to those from the traditional equating methods. These results can be obtained from the first author.

[4] The full set of results for the different simulated sample sizes can be obtained from the first author upon request.