Listening. Learning. Leading.®

Research Report
ETS RR–14-14

Test Score Equating Using Discrete Anchor Items Versus Passage-Based Anchor Items: A Case Study Using *SAT*® Data

Jinghua Liu

Jiyun Zu

Edward Curley

Jill Carey

December 2014

Discover this journal online at
Wiley Online Library
wileyonlinelibrary.com

ETS Research Report Series

EIGNOR EXECUTIVE EDITOR

James Carlson
Principal Psychometrician

ASSOCIATE EDITORS

Beata Beigman Klebanov
Research Scientist

Heather Buzick
Research Scientist

Brent Bridgeman
Distinguished Presidential Appointee

Keelan Evanini
Managing Research Scientist

Marna Golub-Smith
Principal Psychometrician

Shelby Haberman
Distinguished Presidential Appointee

Donald Powers
Managing Principal Research Scientist

Gautam Puhan
Senior Psychometrician

John Sabatini
Managing Principal Research Scientist

Matthias von Davier
Senior Research Director

Rebecca Zwick
Distinguished Presidential Appointee

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ayleen Stellhorn
Editor

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Test Score Equating Using Discrete Anchor Items Versus Passage-Based Anchor Items: A Case Study Using *SAT*® Data

Jinghua Liu,[1,2] Jiyun Zu,[1] Edward Curley,[1] & Jill Carey[1]

1 Educational Testing Service, Princeton, NJ
2 Secondary School Admission Test Board, Princeton, NJ

The purpose of this study is to investigate the impact of discrete anchor items versus passage-based anchor items on observed score equating using empirical data. This study compares an *SAT*® critical reading anchor that contains more discrete items proportionally, compared to the total tests to be equated, to another anchor that contains fewer discrete items and more passage-based items proportionally. Both of these anchors were administered in an SAT administration. The impact of anchor type on equating was evaluated with respect to equating bias. The results clearly reveal that the anchor with more discrete items almost always leads to more accurate equating functions than does the anchor with more passage-based items.

Keywords Anchor test equating; discrete anchor; passage-based anchor; local dependence

doi:10.1002/ets2.12015

In test score equating using data collected from the nonequivalent groups anchor test (NEAT) design, the anchor is used to account for any ability differences between nonequivalent groups taking the new and old forms. It is critical that the two groups of test takers perform in the same way on the anchor as they do on the total test. It is generally considered as a guideline to construct the anchor test by following the same specifications (proportionally) as the total tests—both content and statistical—so that it is a miniature version of the tests to be equated (Angoff, 1984; Dorans, Kubiak, & Melican, 1998; Dorans, Moses, & Eignor, 2010; Kolen & Brennan, 2004). Such a mini-version anchor is expected to do a good job of removing bias and increasing precision in the estimation of the equating function (Holland, Dorans, & Petersen, 2006).

Based on how an item is constructed, a distinction is made between two types of items, *discrete items* and *passage-based items*. Each discrete item stands alone and is unique, whereas a passage-based item is usually administered with other items based on the same stimulus. To equate a test that is composed of both discrete items and passage-based items, it is a common practice to assemble an anchor with both item types that mimics the total test to be equated, based on the mini-version anchor theory. However, a simulation study by Zu and Liu (2010) suggested that a miniature anchor with respect to item types was not ideal for equating tests containing both discrete items and passage-based items. They constructed 20 sets of anchors, which included five different proportions of discrete and passage-based anchor items, two different numbers of items associated with each passage, and two different degrees of local independence among passage-based items. Chained equating (CE) and post-stratification equating (PSE) were conducted, and the equating results were compared and evaluated in terms of equating bias, standard error, and total equating error (i.e., root mean squared error). They found that an anchor with a smaller proportion of passage-based items, fewer items associated with each passage, and/or a smaller degree of local dependence among passage-based items produces smaller equating errors, even if such an anchor may not be a mini-version of the total test in terms of item types. The authors suggested that practitioners consider the possibility of reducing the proportion of passage-based items when they assemble anchors.

The current study can be seen as a follow-up to the Zu and Liu (2010) study. While that study was based on simulations, the current study uses empirical data to investigate the impact of discrete anchor items versus passage-based anchor items on test score equating.

Corresponding author: J. Liu, E-mail: JLiu@ssat.org

Methodology

NEAT Equating

The current study uses operational data from an *SAT*® critical reading section. The SAT reading section comprises 67 multiple-choice items: 19 sentence completion items, which are discrete items, and 48 reading comprehension items, which are passage-based items.

Two forms, Form *X* and Form *Y*, were spiraled and administered in one administration. In the current study, we treated Form *X* as the new form, and Form *Y* as the old form. Each form contains multiple sections, and each section is separately timed. In addition to operational sections in which the items are counted toward the reported scores, there is also one variable section that contains anchor items (or, in some cases, pretest items). The items in the variable section do not count toward test takers' reported scores, so it is an external anchor.

Two anchors were built and administered. Table 1 presents the numbers and proportions of discrete items and passage-based items in each anchor, along with those in the total test. As can be seen from Table 1, one anchor contains 33% discrete items and 67% passage-based items, which is comparable to the total test (28% and 72%, respectively). This anchor is referred to as the *passage-based anchor* for the purpose of the study (although it is more like a miniature anchor). The other anchor is markedly different from the total test in terms of item types: the composition of discrete items and passage-based items is 50% each. This anchor is referred to as the *discrete anchor*. Both anchors were content representative and contained 24 items each with no overlap. The mean and standard deviation of the item difficulty in both anchors were carefully controlled and comparable: the mean of *p* values was .59 for each anchor, and the standard deviations of *p* values were also very similar (0.20 vs. 0.21).

Each form has different subforms. The operational items are identical in different subforms, but the anchor items can vary. As can be seen in Table 2, subform *X1* had operational items and the passage-based anchor, and subform *X2* had identical operational items as *X1* but the discrete anchor. Similarly, subform *Y1* contained operational items and the passage-based anchor, whereas subform *Y2* had identical operational items as *Y1* but the discrete anchor. Hence, subform *X1* could be equated to subform *Y1* via the passage-based anchor, while subform *X2* could be equated to subform *Y*2 using the discrete anchor.

Two classes of observed score equating methods were used: CE and PSE. CE is a chain of two links between the total test and the less reliable anchor test: first, link test *X* to anchor *A* on population *P*, and then link anchor *A* to test *Y* on population *Q*. The two single-group linking functions are then concatenated to map *X* to *Y* through *A*. Two CE methods, the chained linear method and the chained equipercentile method, were conducted in this study. PSE uses the anchor test *A* to post-stratify both *X* and *Y* by first conditioning their distributions on *A* and then reweighting their distributions to estimate the cumulative distribution functions (von Davier, Holland, & Thayer, 2004). The PSE methods used in this study are the Tucker (linear) method and the frequency estimation (nonlinear) method. The weight is proportional to the number of examinees in *P* and *Q*, respectively.

Equating Criterion Based on Equivalent Groups Design

Because forms *X* and *Y* were spiraled and administered in the same SAT administration, the spiraling procedure and the large number of test takers usually yield equivalent groups. Further examination based on the anchors confirmed group equivalence. As can be seen from the descriptive statistics of the raw anchor score[1] presented in Table 3, the new form group who took the passage-based anchor had a mean of 12.10; the old form group who took the same anchor had a very similar mean: 12.13. The standardized mean difference, or the effect size of the mean, was calculated as $\frac{\overline{A}_P - \overline{A}_Q}{SD_{A(P+Q)}}$, where $\overline{A}_P$

Table 1 Test Configuration Total Test, Discrete Anchor, and Passage-Based Anchor

	Total test	Passage-based anchor	Discrete anchor
Total no. of items	67	24	24
No. of discrete items	19	8	12
% of discrete items	28%	33%	50%
No. of passage-based items	48	16	12
% of passage-based items	72%	67%	50%

Table 2 Nonequivalent Groups Anchor Test Design

	Operational items	Passage-based anchor	Discrete anchor
New form *X*			
Subform *X*1	√	√	
Subform *X2*	√		√
Old form *Y*			
Subform *Y*1	√	√	
Subform *Y*2	√		√

Table 3 Raw Score Descriptive Statistics for the Total-Group-to-Total-Group Equating

	Passage-based anchor				Discrete anchor			
	New form sample		Old form sample		New form sample		Old form sample	
Statistics	Test *X*	Anchor	Anchor	Test *Y*	Test *X*	Anchor	Anchor	Test *Y*
Sample size	11,112	11,112	11,187	11,187	11,118	11,118	10,554	10,554
Number of items	67	24	24	67	67	24	24	67
Mean	33.61	12.10	12.13	33.94	33.64	11.99	11.97	33.75
SD	14.12	5.72	5.73	14.05	14.20	5.83	5.73	14.05
Skewness	−0.13	−0.03	−0.02	−0.14	−0.15	−0.02	−0.07	−0.16
Kurtosis	2.37	2.30	2.31	2.37	2.35	2.29	2.34	2.40
Reliability	0.91	0.80	0.79	0.91	0.91	0.80	0.80	0.91
Correlation	0.85		0.84		0.85		0.85	
Standardized mean difference (new/old)			−0.01				0.00	
Ratio of variances (new/old)			1.00				1.04	

is the mean score on the anchor in P, $\overline{A}_Q$ is the mean score on the anchor in Q, and $SD_{A(P+Q)}$ is the standard deviation of the anchor score on the combined population $P+Q$. As can be seen, the standardized mean difference was −0.01 on the passage-based anchor. Similarly, the new/old form groups who took the discrete anchor had very close means (11.99 vs. 11.97) and a standardized mean difference of 0. Hence, the groups taking forms *X* and *Y* were deemed equivalent.

We then equated form *X* to *Y* via equivalent groups design with no anchor.[2] Both nonlinear (equipercentile) and linear (mean sigma) equating functions were conducted. The difference between the linear equating function and nonlinear equating function was trivial across the entire scale range. Hence, the linear equating function was deemed more appropriate and was used as the equating criterion in this study.

Equating Samples

Much research evidence has accumulated and suggests that when an anchor is used to equate test scores, the ability differences in the groups taking the old and new forms affect the quality of equating (Cook & Petersen, 1987; Kolen, 1990). Since our total group equating samples were fairly similar, we identified several subgroups (SGs) with different abilities based on the examinees' demographic information to simulate equating scenarios with different levels of group difference.

The SGs are referred to as SG1 to SG5. These SGs are mutually exclusive but are not exhaustive subpopulations. We then constructed four pairs of equating samples with varied ability differences. Two sets of pairs had similar ability levels: (a) total group and total group and (b) SG1 and SG2. The other two sets of pairs had large ability differences: (c) SG3 and SG5 and (d) SG4 and SG5. In each pair, we used one member as the new form sample and the other as the old form sample. For example, test takers belonging to SG3 and taking form *X* and those belonging to SG5 and taking form *Y* were used to equate *X* to *Y* to emulate an equating with different samples. Note that there was one pair of SG1 – SG2 who took the new/old forms with the passage-based anchor; there was another pair of SG1 – SG2 who took the new/old forms with the discrete anchor. Hence, for each degree of ability difference in the equating samples, we performed equating twice, once using the passage-based anchor (i.e., subform *Y1* to subform *X1*) and once using the discrete anchor (i.e., subform *Y2* to subform *X2*).

Table 4 Raw Score Descriptive Statistics for the SG1-to-SG2 Equating

	Passage-based anchor				Discrete anchor			
	New form sample		Old form sample		New form sample		Old form sample	
Statistics	Test *X*	Anchor	Anchor	Test *Y*	Test *X*	Anchor	Anchor	Test *Y*
Sample size	6,336	6,336	4,892	4,892	6,341	6,341	4,625	4,625
Number of items	67	24	24	67	67	24	24	67
Mean	32.66	11.88	12.40	35.27	32.94	11.89	12.15	34.89
SD	14.15	5.72	5.72	14.06	14.19	5.84	5.80	14.18
Skewness	−0.07	0.00	−0.08	−0.26	−0.11	0.01	−0.09	−0.21
Kurtosis	2.36	2.34	2.32	2.46	2.34	2.26	2.33	2.40
Reliability	0.91	0.80	0.79	0.91	0.91	0.80	0.80	0.91
Correlation	0.85		0.85		0.85		0.85	
Standardized mean difference (new/old)			−0.09				−0.04	
Ratio of variances (new/old)			1.00				1.01	

Table 5 Raw Score Descriptive Statistics for the SG3-to-SG5 Equating

	Passage-based anchor				Discrete anchor			
	New form sample		Old form sample		New form sample		Old form sample	
Statistics	Test *X*	Anchor A1	Anchor A1	Test *Y*	Test *X*	Anchor A2	Anchor A2	Test *Y*
Sample size	1,760	1,760	6,003	6,003	1,736	1,736	5,721	5,721
Number of items	67	24	24	67	67	24	24	67
Mean	27.66	10.07	13.18	36.86	26.92	9.42	13.08	36.47
SD	13.88	5.50	5.25	12.63	13.54	5.53	5.26	12.53
Skewness	0.15	0.18	−0.08	−0.19	0.17	0.25	−0.12	−0.20
Kurtosis	2.38	2.40	2.43	2.56	2.26	2.47	2.44	2.57
Reliability	0.91	0.78	0.76	0.89	0.90	0.78	0.77	0.90
Correlation	0.84		0.82		0.84		0.83	
Standardized mean difference (new/old)			−0.57				−0.66	
Ratio of variances (new/old)			1.10				1.10	

The raw score descriptive statistics for each pair of equating samples are presented in Tables 3 – 6 for total-to-total, SG1 to-SG2, SG3-to-SG5, and SG4-to-SG5, respectively. Tables 3 – 6 also list the standardized mean differences on the anchor tests.

As discussed above, data in Table 3 show that the two total groups were very similar, with the standardized mean difference on both anchors close to zero. SG1 and SG2 (Table 4) showed a little bit larger difference in ability level (the standardized mean difference was −0.09 on the passage-based anchor and −0.04 on the discrete anchor) compared to the pair of total groups, but the differences were still considered small. In contrast, the SG3 and SG5 were quite different (Table 5): the standardized mean differences were −0.57 and −0.66 on the passage-based anchor and the discrete anchor, respectively; SG3 was much less able than SG5. Finally, the SG4-to-SG5 comparison shown in Table 6 exhibits the largest ability difference, with the standardized mean difference approaching −0.80 on both anchors. SG4 was the least able group, very much less able than SG5. Tables 3 – 6 also present the correlation between the anchor and the total test for each equating sample. The results suggest that the correlations based on the discrete anchor were very similar to those based on the passage-based anchor.

Discrepancy Index: Bias

Although bias and standard error of equating (SEE) are usually used in studies that evaluate equating accuracy, previous studies have found that the anchor type does not seem to have much impact on SEE (Liu, Sinharay, Holland, Feigenbaum, & Curley, 2011; Yi, 2009; Zu & Liu, 2010). The anchor type effect on the root mean square error (RMSE) is predominantly determined by bias. We also observed in the current study that there was little difference in the SEE of passage-based and discrete anchors. Hence, we report only the equating bias.

Table 6 Raw Score Descriptive Statistics for the SG4-to-SG5 Equating

	Passage-based anchor				Discrete anchor			
	New form sample		Old form sample		New form sample		Old form sample	
Statistics	Test *X*	Anchor A1	Anchor A1	Test *Y*	Test *X*	Anchor A2	Anchor A2	Test *Y*
Sample size	1,174	1,174	6,003	6,003	1,219	1,219	5,721	5,721
Number of items	67	24	24	67	67	24	24	67
Mean	25.19	8.97	13.18	36.86	24.79	8.67	13.08	36.47
SD	13.20	5.37	5.25	12.63	13.48	5.42	5.26	12.53
Skewness	0.29	0.39	−0.08	−0.19	0.30	0.41	−0.12	−0.20
Kurtosis	2.45	2.63	2.43	2.56	2.54	2.70	2.44	2.57
Reliability	0.89	0.76	0.76	0.89	0.90	0.77	0.77	0.90
Correlation	0.82		0.82		0.83		0.83	
Standardized mean difference (new/old)			−0.77				−0.79	
Ratio of variances (new/old)			1.04				1.06	

The bias at each score point, or the conditional bias, can be calculated by

$$\text{Bias}\left[\hat{e}_Y\left(x_i\right)\right] = \hat{e}_Y\left(x_i\right) - e\left(x\right), \tag{1}$$

where $\hat{e}_Y(x_i)$ is the sample equating function at score point x_i derived from the anchor test equating, and $e(x)$ is the criterion equating function based on equivalent groups equating.

To summarize the differences over all score points, we calculate the weighted absolute bias given by

$$\text{Weighted absolute bias}\left[\hat{e}_Y\left(x\right)\right] = \frac{1}{N}\sum f_{x_i}\left|\text{Bias}\left[\hat{e}_Y\left(x_i\right)\right]\right|, \tag{2}$$

where N is the total number of test takers in the new form equating sample, and f_{x_i} is the frequency at score point x_i in the new form group.

In equating studies, it is usually useful to have a threshold to evaluate the magnitude of an equating error. Some authors, such as Dorans and Feigenbaum (1994), have elected to use 0.5 on the raw score scale as a rough guideline for an acceptable level of equating error. We use the same guideline in this study.

Results

In this section, we first compare the conditional equating bias and then summarize the effects of anchor types, ability differences, and equating methods.

Conditional Equating Bias

Figure 1 presents the bias results based on frequency estimation equating. In the total-to-total equating, the bias is trivial. The two bias curves intertwine with the zero line throughout the entire scale. In the SG1-to-SG2 equating, the magnitude of the bias increases, yet the two bias curves are again very close. The discrete anchor performs slightly better in the lower to middle of the score range (approximately from raw scores 20 to 40), whereas the passage-based anchor performs a little better at the ends of the scale. When the ability difference increases, both anchors produce much larger bias, and the discrepancy between the two different anchors increases as well. The discrete anchor produces smaller bias in the SG3-to-SG5 equating, except at the high end of the scale, and it also does better than or as well as the passage-based anchor in the SG4-to-SG5 equating, except at the very low end of the scale.

Results based on the Tucker method are presented in Figure 2. The basic pattern is very similar to what is observed in Figure 1. Both anchors produce accurate results in the total-to-total equating. When the two equating samples are slightly different, as in the SG1-to-SG2 equating, the equating bias increases (around 1.5), yet the two anchors still produce similar results. In both equating scenarios, although close, the discrete anchor does result in slightly smaller bias below raw score 30. However, when the two groups are different, the two anchors produce quite different results: the discrete anchor does better across virtually the entire scale range in both the SG3-to-SG5 equating and the SG4-to-SG5 equating.

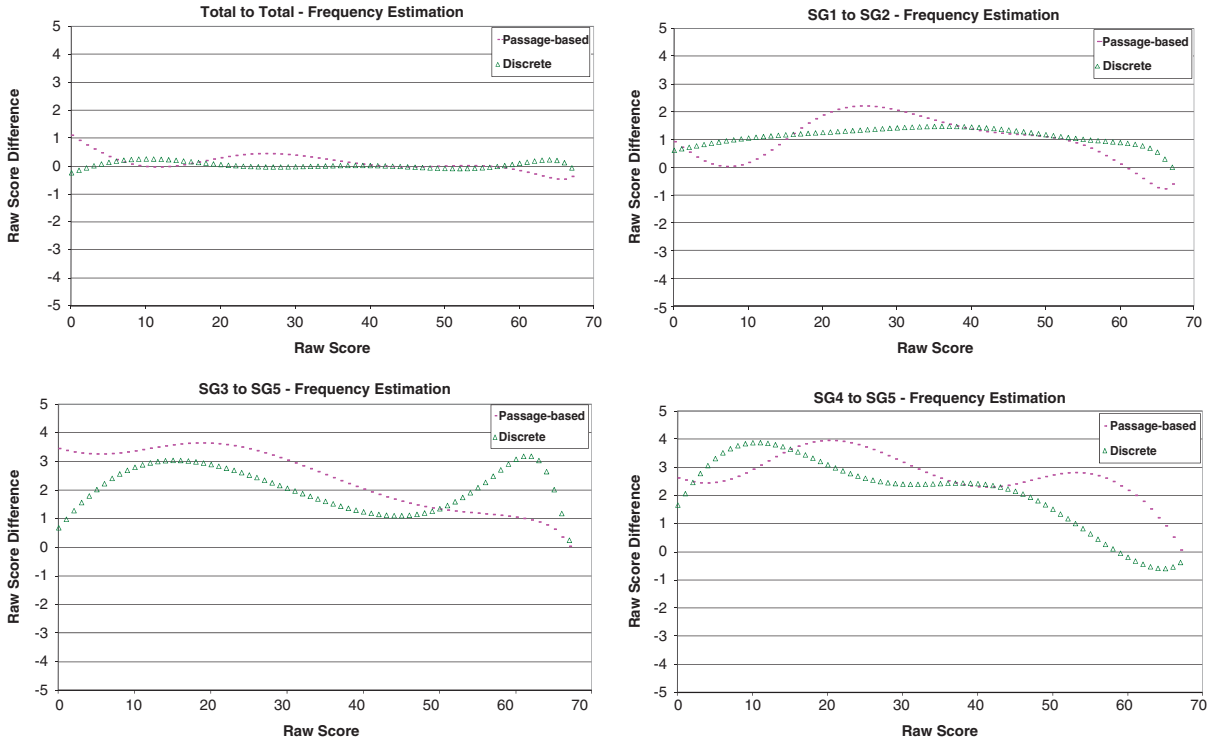

Figure 1 Bias based on frequency estimation equating.

Figure 3 shows the bias plots for chained equipercentile equating. In the total-to-total equating, the two bias curves are close to each other across the entire scale. The discrepancy between the two bias curves somewhat increases in the SG1-to-SG2 equating. The discrete anchor performs better in the lower half of the score range, whereas the passage-based anchor performs better elsewhere. When the equating samples are very different, the discrete anchor in general produces smaller bias than the passage-based anchor, with just a couple of exceptions: around the high end (above raw score 50) in the SG3-to-SG5 equating and near the low end (below raw score 20) in the SG4-to-SG5 equating.

Chained linear results are plotted in Figure 4. The pattern is somewhat similar to what is observed in Figure 2 (results yielded by Tucker method). When the ability differences between the two equating samples are small, the difference of bias is trivial. When the ability differences are large, however, the discrete anchor performs better than the passage-based anchor across almost the entire score range.

Across all the equating methods, the magnitude of the bias is less than 0.5 in the total-to-total equatings, and lingering between 0 and 2 in the SG1-to-SG2 equatings. The bias increases in the SG3-to-SG5 and SG4-to-SG5 equatings, exceeding 2 (CE method) or even approaching 5 (PSE method).

Anchor Type Effects

Table 7 summarizes the weighted absolute bias. A comparison of the overall bias between the discrete anchor and the passage-based anchor is graphically presented in Figure 5. The pattern is clear: the discrete anchor consistently produces smaller overall bias than the passage-based anchor. This finding is consistent with the results from Zu and Liu (2010). The only exception is in the SG1-to-SG2 equating using the chained linear method, where the discrete anchor and the passage-based anchor result in pretty much the same amount of bias.

Group Ability Difference Effects

Group ability difference effects on bias are also reflected by Figure 5. When the group ability difference is near zero, equating is essentially unbiased. When the ability differences increase, the bias increases as well. Two exceptions are observed, both of which happen with the CE methods. One exception is that the SG1-to-SG2 equating produces larger equating bias

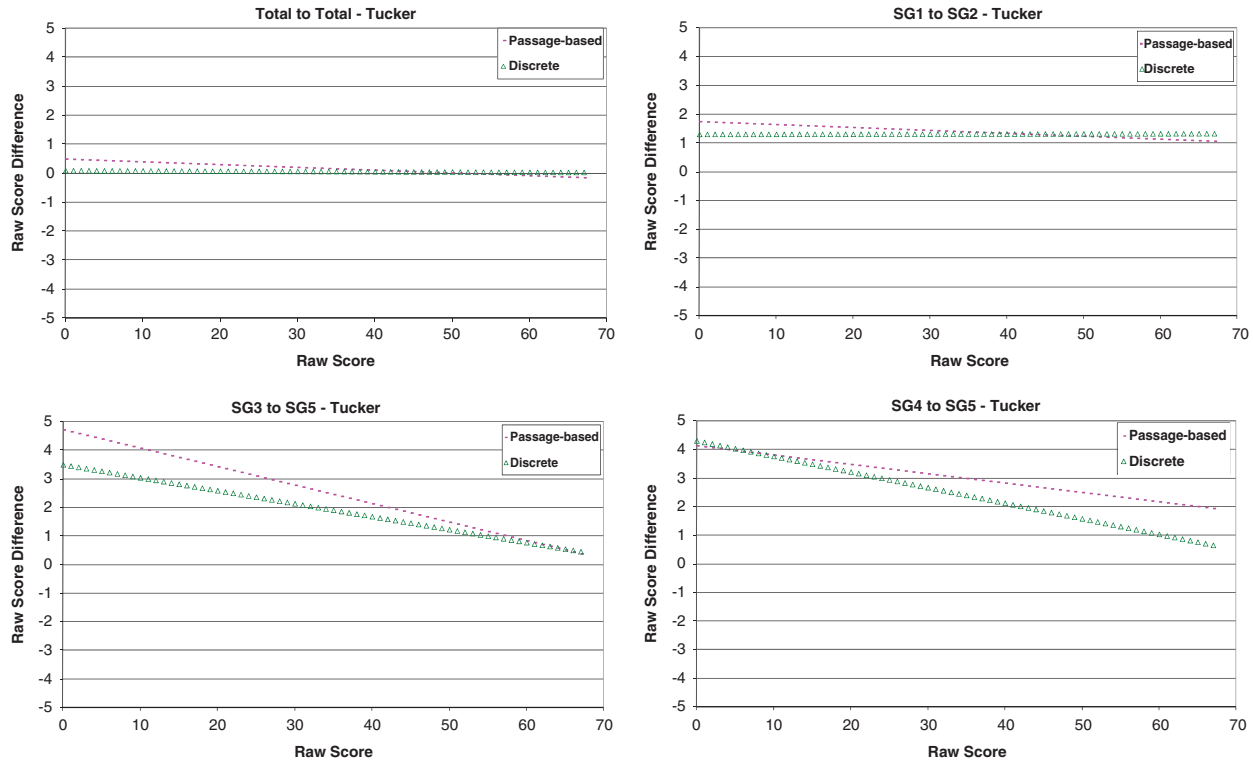

Figure 2 Bias based on Tucker equating.

than either the SG3-to-SG5 or SG4-to-SG5 equating via discrete anchor, even though the SG1-to-SG2 has much smaller ability difference between the new- and old-form equating samples. The other exception is that the SG4-to-SG5 equating yields smaller bias than the SG3-to-SG5 equating using the passage-based anchor, although the latter has a relatively smaller ability difference than does the former.

Equating Method Effects

As can be seen in both Figure 5 and Table 7, PSE and CE methods produce very similar equating bias when the group ability difference is small, but when the ability difference increases, CE produces smaller equating bias. This finding is consistent with previous research (Wang, Lee, Brennan, & Kolen, 2008; Sinharay & Holland, 2007). This result is expected. PSE methods require population invariance assumptions: the conditional distribution of X given anchor A is the same in populations P and Q, and the conditional distribution of Y given anchor A is the same in populations P and Q. The more equivalent P and Q are, the more likely these two invariance assumptions will hold. As the differences of the groups get larger and the differences of the covariate get larger, the tendency to regress the solution toward equivalent groups grows in strength (Holland & Dorans, 2006; von Davier et al., 2004).

Interaction Among Anchor Types, Ability Differences, and Equating Methods

Interaction among anchor types, ability differences, and equating methods is observed. When the ability difference is minimal, anchor types have trivial influence on bias with the discrete anchor producing slightly smaller bias; equating methods do not seem to have too much impact on bias. When the ability difference is substantial, the discrete anchor produces much smaller bias than the passage-based anchor; PSE methods produce larger equating bias than the CE methods.

Discussion

The NEAT design is often employed for test score equating. In the NEAT design, an anchor test plays a crucial role in accounting for ability differences between the groups of test takers who are administered different forms. It is usually

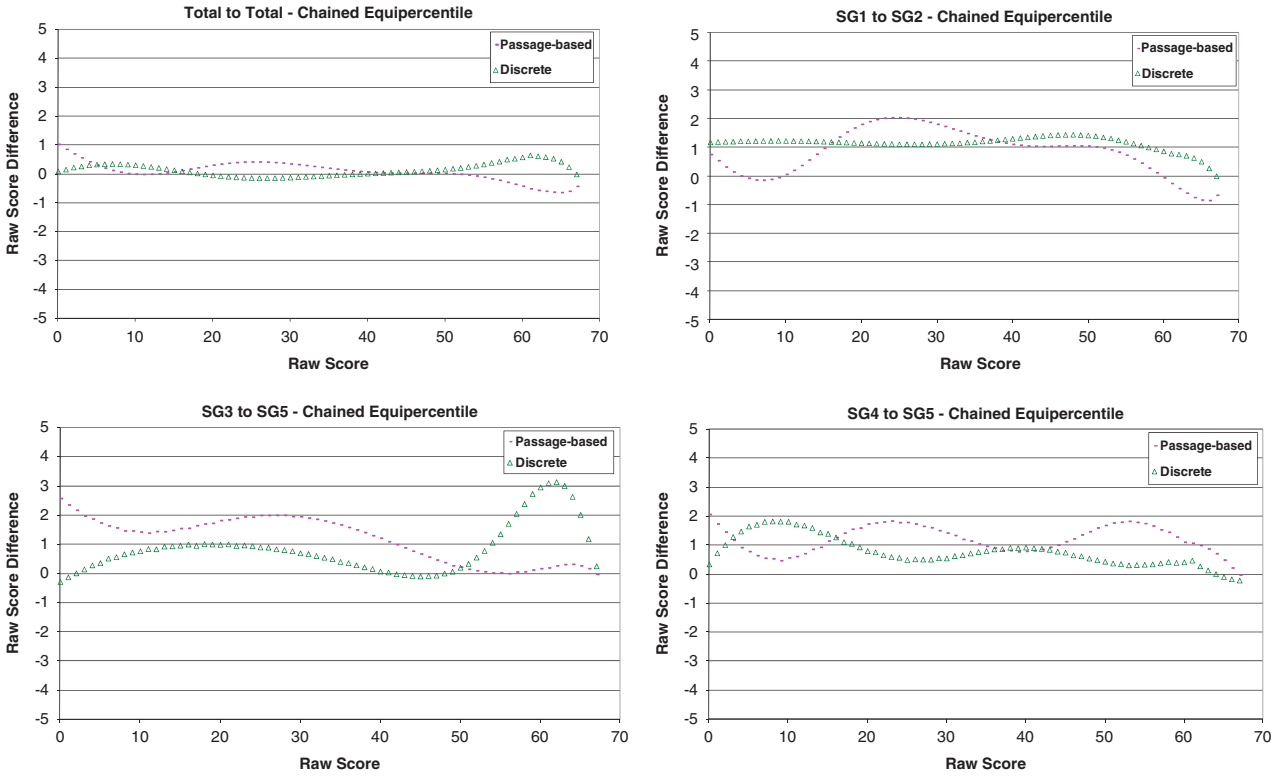

Figure 3 Bias based on chained equipercentile equating.

considered good practice to assemble an anchor test following the same specifications (proportionally) as the tests to be equated. Such an anchor is also referred to as a miniature of the total tests.

This study explores an anchor that contains more discrete anchor items proportionally compared to the tests to be equated. This anchor, in contrast to another anchor that contains a similar proportion of discrete items and passage-based items as the total tests, were both given in an SAT administration. The impact of anchor type on equating was evaluated with respect to equating bias. The results clearly reveal that the anchor with more discrete items almost always leads to more accurate equating functions than the anchor with more passage-based items, even though the latter was a miniature of the total tests. Hence, these empirical results confirm the findings from Zu and Liu's (2010) simulation study.

The main reason for the larger bias via the passage-based anchor than via the discrete anchor could be that passage-based items tend to lower the reliability of a test. This is because items associated with the same passage are often locally dependent, perhaps due to different degrees of familiarity with the subject matter of the passage. While alternate forms of the same test have passages with a range of topics, the reliability (i.e., the correlation of test scores from parallel forms) is often smaller than that of a similar test composed of all discrete items that are designed to measure the same latent ability (Lawrence, 1995; Wainer & Thissen, 1998). The lower reliability could reduce the correlation between the total test and the anchor, which further diminishes the quality of observed score equating. As Zu and Liu (2010) pointed out, "the existence of passage-specific constructs reduces the proportion of the variance due to latent ability, thus reduces the correlation between the anchor and the total tests, which leads to larger equating errors" (p. 410). Therefore, our recommendation is to reduce the proportion of passage-based items in the anchor and increase the proportion of discrete items if possible, given other program-specific issues and fairness-related concerns.

Although our current study did not reveal large differences of the reliabilities by anchor type (the reliabilities based on the discrete anchor were either almost identical or slightly higher than those based on the passage-based anchor, as shown in Tables 3 – 6), reliability may still be a factor resulting in the equating bias differences. In the current study, we calculated the reliabilities using a traditional method: Cronbach's alpha, which treats each item as an independent item, regardless

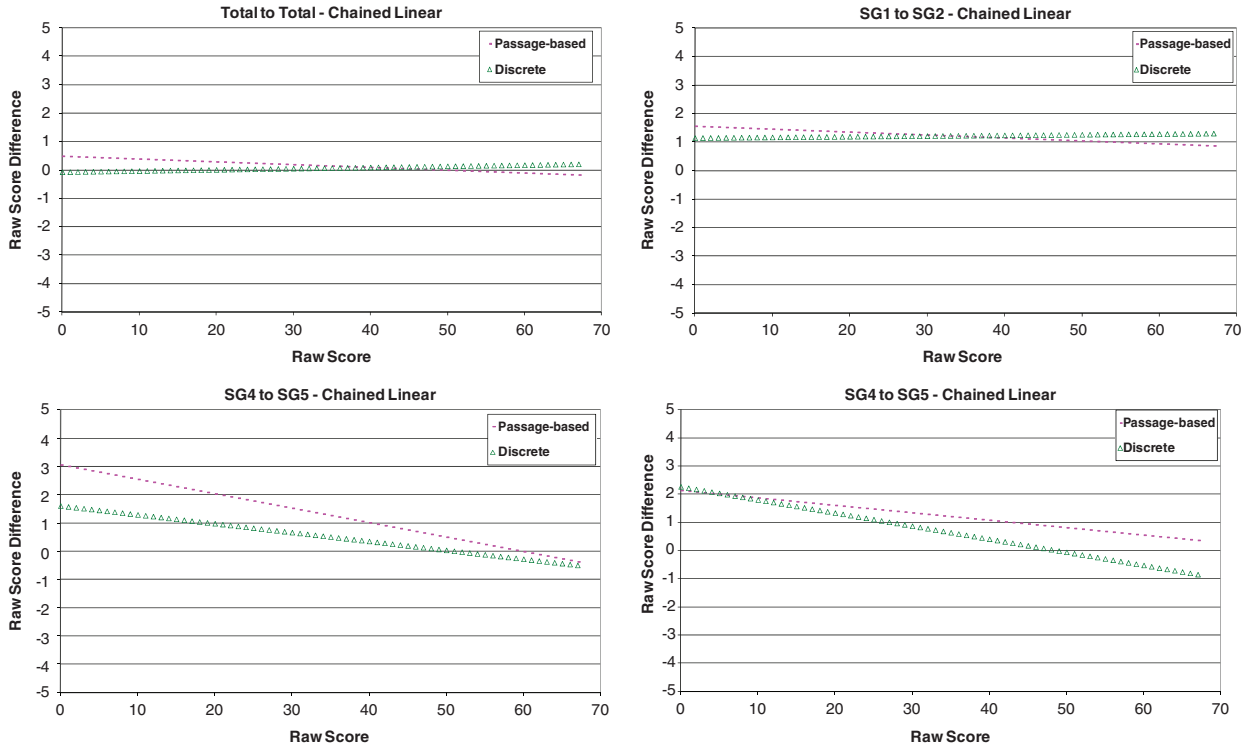

Figure 4 Bias based on chained linear equating.

Table 7 Summary of Weighted Absolute Bias

	Tucker		Frequency estimation		Chained linear		Chained equipercentile	
	Passage	Discrete	Passage	Discrete	Passage	Discrete	Passage	Discrete
Total – total	0.18	0.06	0.19	0.06	0.17	0.08	0.20	0.13
SG1 – SG2	1.42	1.32	1.42	1.29	1.23	1.23	1.25	1.22
SG3 – SG5	2.95	2.25	2.89	2.17	1.65	0.76	1.50	0.66
SG4 – SG5	3.32	2.94	3.15	2.82	1.47	1.12	1.24	0.96

of whether it is a discrete item or an item based on a passage. Sireci, Thissen, and Wainer (1991) found that failing to take into account the item dependencies caused by a common passage would be likely to yield a 10 – 15% overestimate of reliability. In future research, it might be worthwhile to collapse the passage scores for each passage, and treat all the items based on the same passage as one *super item* and then estimate the reliability. This should yield more accurate reliability estimates.

In addition to the effect of anchor type, the effect of equating method and ability differences were also investigated in this study. The anchor type does not seem to interact with equating method. The discrete anchor produced more accurate equatings than the passage-based anchor, whether it was via PSE or CE. On the other hand, anchor type seems to interact somewhat with ability difference: when the two groups were similar, the discrete anchor produced slightly better results; when the two groups were substantially different, the discrete produced much better results.

In this study, it was observed that group differences in ability have a substantial impact on equating bias. In general, the larger the ability differences between the two equating samples, the larger the bias, especially with the PSE methods. This finding is consistent with other research findings (Kolen, 1990; MacCann, 1990; Sinharay & Holland, 2007; Wang et al., 2008; Zu & Liu, 2010). What is interesting is the interaction between equating methods and ability differences. PSE methods produced equating bias in a very consistent pattern: the larger the ability difference, the larger the bias, regardless of whether a discrete anchor or a passage-based anchor was used. CE methods, on the other hand, exhibited a different pattern. While the overall trend of bias was still increasing along with the increase of the ability difference, the

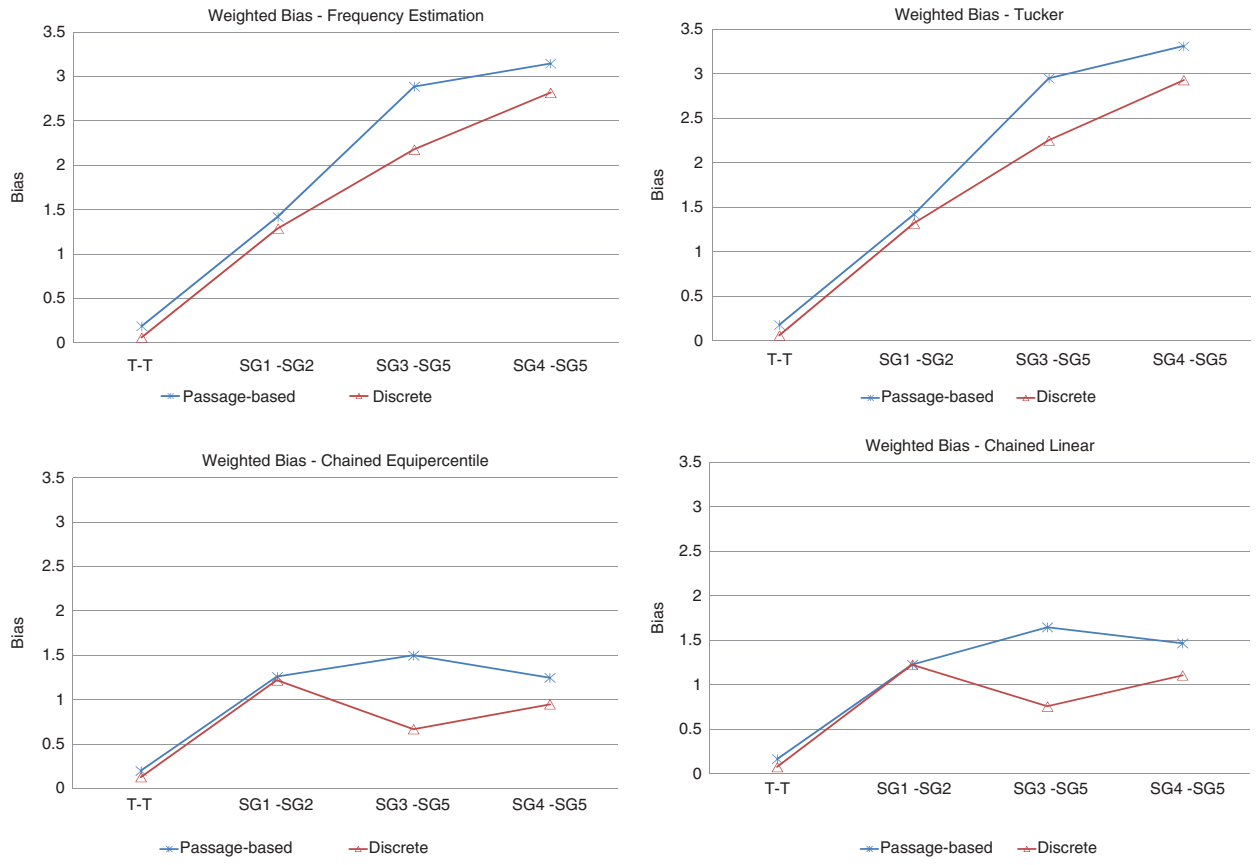

Figure 5 Comparison of weighted bias between anchor types based on observed score equating.

greatest amount of bias was not necessarily associated with the pair of equating samples that were the most different. For example, the largest bias via the discrete anchor was produced in the SG1-to-SG2 equating, and the largest bias via the passage-based anchor occurred in the SG3-to-SG5 equating (even though the SG4-to-SG5 equating exhibited the largest ability difference). It seems that CE is not very sensitive, or not sensitive systematically, to the ability differences between the equating samples. A possible reason is that CE is essentially a scaling method. Instead of treating the anchor as a covariate, CE just scales test X to anchor A on population P, and then scales anchor A to test Y on population Q. The difference between populations P and Q does not matter too much because X and Y are treated as if they had been scaled together on the same population.

The use of discrete anchor items has practical benefits. An anchor with a relatively smaller proportion of passage-based items may potentially reduce some security problems. For example, if test takers remember a passage and then post it online, many items could be compromised at once, and thus the equating could be contaminated.

From the perspective of item/test construction, it is also easier to build a discrete anchor test that meets statistical specifications. This is because it is more difficult to write very easy or very hard passage-based items than it is to write very easy or very hard discrete items. Most passage-based items are of middle difficulty, which can make it challenging to meet statistical specifications. Moreover, when assembling an anchor with more discrete items, it is easier to swap the discrete items than it is to swap passage-based items if there is a need to increase or decrease the mean or standard deviation of the difficulty of an anchor test.

Further investigation of this topic could involve the analysis of more operational data sets from different testing programs, especially testing programs in which discrete items and passage-based items measure different dimensions. Theoretically, it is possible that, for example, a new form sample of test takers performs better than the old form sample of test takers on discrete items but worse on passage-based items, if discrete items and passage-based items measure different dimensions. If the anchor consists of predominantly discrete items and is used to adjust for form difficulty differences, the ability of new form test takers is likely to be overestimated, which could cause the new form difficulty to be overestimated.

The new form test takers may receive unfairly higher scaled scores. We need to be cautious not to overgeneralize the results of the present study.

It may be useful to compare discrete and passage-based anchors with respect to other equating criteria, such as the same distributions property (Kolen & Brennan, 2004) and the first- and second-order equity property (e.g., Tong & Kolen, 2005). Further, note that only observed score equating methods were used in this study. Whether or not the discrete anchor will be similarly robust using item response theory equating methods is also worthy of investigation.

Acknowledgments

The authors thank Neil Dorans for the inspiration to conduct this research. The authors also thank Rebecca Zwick, Neil Dorans, Gautam Puhan, and the three anonymous reviewers for their advice. Any opinions expressed in this publication are those of the authors and not necessarily those of Secondary School Admission Test Board or Educational Testing Service.

Notes

1 What we refer to as a *raw score* is a formula score, where for a multiple-choice item with k number of answer choices, an examinee receives a score of 1 for a correct answer, 0 for a nonresponse, and $-1/(k-1)$ for a wrong answer.

2 Because subforms $X1$ and $X2$ contained identical operational items, and subforms $Y1$ and $Y2$ contained identical operational items, we combined the data of $X1$ and $X2$, and combined data of $Y1$ and $Y2$ to equate X to Y.

References

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.

Cook, L. L. & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement*, *11*(3), 225–244.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel equating method of equating*. New York, NY: Springer-Verlag.

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

Dorans, N. J., Kubiak, A., & Melican, G. J. (1998). *Guidelines for selection of embedded common items for score equating* (ETS Statistical Report No. SR-98-02). Princeton, NJ: Educational Testing Service.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). *Principles and practices of test score equating* (Research Report No. RR-10-29). Princeton, NJ: Educational Testing Service.

Holland, P. W. & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: American Council on Education/Praeger.

Holland, P. W., Dorans, N. J., & Petersen, N. S. (2006). Equating test scores. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (pp. 169–203). Amsterdam, Netherlands: Elsevier.

Kolen, M. J. (1990). Does matching in equating work? A discussion. *Applied Measurement in Education*, *3*(1), 97–104.

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

Lawrence, I. M. (1995). *Estimating reliability for tests composed of item sets* (Research Report No. RR-95-18). Princeton, NJ: Educational Testing Service.

Liu, J., Sinharay, S., Holland, P. W., Feigenbaum, M., & Curley, E. (2011). Observed score equating using a mini-version anchor and an anchor with less spread of difficulty: A comparison study. *Educational and Psychological Measurement*, *71*(2), 346–361.

MacCann, R. G. (1990). Derivations of observed score equating methods that cater to populations differing in ability. *Journal of Educational Statistics*, *15*(2), 146–170.

Sinharay, S. & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement*, *44*(3), 249–275.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 237–247.

Tong, Y. & Kolen, M. J. (2005). Assessing equating results on different equating criteria. *Applied Psychological Measurement*, *29*(6), 418–432.

Wainer, H., & Thissen D. (1998). *How is reliability related to the quality of test scores? What is the effect of local dependence on reliability?* (Research Report No. RR-98-1). Princeton, NJ: Educational Testing Service.

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

Yi, H. S. (2009, April). *Evaluating the performance of non-equivalent groups with anchor test equating under various conditions of anchor test construction.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Zu, J. & Liu, J. (2010). Observed score equating using discrete and passage-based anchor items. *Journal of Educational Measurement*, *47*(4), 395–412.

Suggested citation:

Liu, J., Zu, J., Curley, E., & Carey, J. (2014). *Test score equating using discrete anchor items versus passage-based anchor items* (Research Report No. RR-14-14). Princeton, NJ: Educational Testing Service. doi: 10.1002/ets2.12015

Action Editor: Rebecca Zwick

Reviewers: Neil Dorans and Gautam Puhan

ETS, the ETS logo, and LISTENING. LEARNING. LEADING. are registered trademarks of Educational Testing Service (ETS). SAT is a registered trademark of the College Board. All other trademarks are property of their respective owners.

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/