Research Report

Comparison of the Effects of Discrete Anchor Items and Passage-Based Anchor Items on Observed-Score Equating Results

Jiyun Zu

Jinghua Liu

December 2009

ETS RR-09-44

Comparison of the Effects of Discrete Anchor Items and Passage-Based Anchor Items on Observed-Score Equating Results

Jiyun Zu and Jinghua Liu
ETS, Princeton, New Jersey

December 2009

Abstract

Equating of tests composed of both discrete and passage-based items using the nonequivalent groups with anchor test (NEAT) design is popular in practice. This study investigated the impact of discrete anchor items and passage-based anchor items on observed score equating via simulation. Results suggested that an anchor with a larger proportion of passage-based items and/or a larger degree of local dependence among passage-based items produces larger equating errors, especially when group ability differences are not minimal. Our findings challenge the common belief that an anchor should be a miniature version of the tests to be equated.

Key words: Passage-based anchor items, discrete anchor items, testlet model, the nonequivalent groups with anchor test (NEAT) design, chain equating method, post-stratification equating method

Acknowledgments

The authors gratefully acknowledge the help of Tim Moses with SAS programming. The authors thank Sandip Sinharay and Neil J. Dorans for their invaluable advice. Any opinions expressed in this study are those of the authors and not necessarily of ETS.

Table of Contents

Page

Methodology 2

Test and the Equating Design 2

Data Generation 3

Factors Studied in the Simulation 4

Simulation Steps 5

Evaluation Criteria 6

Results 9

Effects on Equating Bias 9

Effects on Standard Error of Equating 13

Effects on Root Mean Squared Error 13

Conclusions and Discussion 16

References 23

List of Tables

Page

Table 1 The Equating Design 3

Table 2 Weighted Average of Absolute Bias (×100) Under Different Conditions 9

Table 3 Weighted Average of Standard Error of Equating (×100) Under Different Conditions 13

Table 4 Weighted Average of Root Mean Squared Error (×100) Under Different Conditions 17

List of Figures

Page

Figure 1. Empirical distributions of *X* under different conditions. 7

Figure 2. Weighted average of absolute bias (×100) under different conditions. 10

Figure 3. Conditional bias for 0% passage-based anchor items. 11

Figure 4. Conditional bias for 50% passage-based anchor items. 12

Figure 5. Weighted average of standard error of equating (×100) under different conditions. 14

Figure 6. Conditional standard error of equating for 0% passage-based anchor items. 15

Figure 7. Conditional standard error of equating for 50% passage-based anchor items. 16

Figure 8. Weighted average of root mean squared error (×100) under different conditions. 18

Figure 9. Conditional root mean squared error for 0% passage-based anchor items. 19

Figure 10. Conditional root mean squared error for 50% passage-based anchor items. 20

The nonequivalent groups with anchor test (NEAT) design is often employed for test score equating. In the NEAT design, population *P* takes the new form *X*, population *Q* takes the old form *Y*, and both populations take the same anchor test *V*. The anchor test is used to estimate differential ability by quantifying the difference in ability between the samples of test takers from *P* and *Q*. 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 (Livingston, 2004).

It is commonly advised that an anchor should be a miniature version of the tests to be equated. Kolen and Brennan (2004) suggested that anchor tests should be built to the same specifications as the total test, in both content and statistical characteristics, in order to reflect the group differences accurately (pp. 9, 271). Livingston (2004) recommended choosing common items that resemble the full test in content and format, and represent the full range of difficulty (p. 38).

Based on how an item is constructed, we can divide items into two categories: *discrete items* versus *passage-based items*. Each discrete item is based on a unique item stem (e.g., a sentence completion question), whereas a passage-based item is based on a single stimulus and is meant to be administered together with other items based on the same stimulus (i.e., a set of reading comprehension items that are related to a common reading passage).

Based on the miniature version anchor theory, the anchor items should resemble the full test in content and format. Thus, if the two tests to be equated consist of both discrete items and passage-based items, the anchor should contain both item types as well. However, such an anchor may denigrate equating performance due to two reasons. First, passage-based items tend to be locally dependent because responses to different items are determined by the understanding of the same stimulus. Research has found that local dependence lowers the reliability of the test (Lawrence, 1995; Wainer & Thissen, 1998), which could reduce the correlation between the total test and the anchor, and a high total-anchor correlation is one of the keys for a successful observed score equating. Second, practitioners often encounter security problems when they use anchors containing passage-based items. A reading passage is much easier to memorize than a collection of discrete items. Once a test taker memorizes the passage, or even worse, posts it on a Web site, all the items based on this passage will be compromised; consequently, the equating results will be contaminated. Hence, our question is: what is the ideal construction of an anchor with respect to item types when tests to be equated contain both discrete items and passage-based items? Is it really necessary to follow the miniature anchor theory and to build an anchor with the same proportion of passage-based items as in the total tests to be equated?

Sinharay and Holland (2006, 2007) questioned the restriction that an anchor test needs to have the same spread of difficulty as the tests to be equated. They studied the correlation between the scores of a total test and anchors with different standard deviations of the difficulty parameter. The results showed that anchors with a spread of item difficulties less than that of the total test consistently had higher correlations than miniature anchor tests (Sinharay & Holland, 2006). The authors concluded that it might be not optimal and might be too restrictive to require an anchor test to mimic the statistical characteristics of the total test. In a follow-up study, Sinharay and Holland (2007) compared the equating performance of anchors with different spreads of item difficulties. Their results suggested that "content-representative anchor tests with item difficulties that are centered appropriately but have less spread than those of total tests perform as well as mini tests in equating" (Sinharay & Holland, 2007, p. 272). Another study by Liu, Sinharay, Holland, Feigenbaum, and Curley (2009) also demonstrated that a miniature anchor test that is both content- and statistical-representative does not necessarily perform better than a nonminiature version anchor (where the nonminiature version anchor is content-representative but not statistical-representative) in terms of equating results. Although not directly related to the current study, these research findings suggest that the miniature anchor theory may not be a panacea.

The purpose of this study is to investigate the impact of discrete anchor items versus passage-based anchor items on observed score equating via simulation. We hope to provide some useful information about anchor design for practitioners.

Methodology

Test and the Equating Design

In this study, test X (the new form) is equated to test Y (the old form) through an external anchor V under a NEAT design. Both tests X and Y have 80 items, 50% (40) of which are discrete items, and the other 50% (40) are passage-based items. The external anchor V contains 40 items, with different proportions of discrete items and passage-based items. The number of items associated with a reading passage is fixed at 10. Hence, an anchor can be composed of 0 passage-based items and 40 discrete items, or 40 passage-based items based on 4 passages and 0 discrete items, or other compositions. We design tests X and V to have the same average difficulty, while *Y* is harder. A sample of size 4,000 from population *P* is taking test *X* and the external anchor *V*. A sample of the same size from population *Q* is taking test *Y* and *V*. The design table is shown in Table 1.

Table 1

The Equating Design

Population	New Form X	Old Form Y	External Anchor V
P	✓		✓
Q		✓	✓

Data Generation

Data generation for discrete items. We simulate the discrete items from the three-parameter uni-dimensional item response theory (IRT) model, where the probability that the *ith* examinee correctly answers an item *j* is expressed as:

$$P(X_{ij} = 1) = c_j + (1 - c_j)\frac{e^{t_{ij}}}{1 + e^{t_{ij}}}, \tag{1}$$

with

$$t_{ij} = a_j(\theta_i - b_j). \tag{2}$$

Parameter θ_i is the latent ability the test is designed to measure for examinee *i*, a_j is the discrimination parameter, b_j is the difficulty parameter, and c_j is the guessing parameter for item *j*.

Data generation for passage-based items. To represent the local dependence among passage-based items, the 3PL testlet model (Wainer, Bradlow, & Wang, 2007, p. 135) was used for data generation. A testlet is defined as an aggregation of items that are based on a single stimulus and that are meant to be administered together (Wainer et al.). A set of passage-based items associated with one reading passage is a testlet. In this model, the probability of the *ith* examinee correctly answering the *jth* item in the *dth* passage is expressed by replacing t_{ij} in (2) with

$$t_{ij} = a_j(\theta_i - b_j - \gamma_{id(j)}) . \tag{3}$$

$\gamma_{id(j)}$ is the testlet parameter for examinee *i* with passage $d(j)$. It represents the dependence among items associated with one reading passage, but is not explained by the ability. The subscript *j* is put in parentheses to indicate that for the same examinee *i*, each item *j* nested within the same passage *d* takes the same value of the testlet parameter. The testlet model is a multidimensional IRT model. For a test containing *D* passages, the testlet model has *D* + 1 dimensions: one general dimension of the latent ability θ and *D* passage-specific dimensions. It has also been shown that the testlet model is equivalent to a second-order factor model for ordered categorical data (Rijmen, 2009). Usually, $\gamma_{id(j)}$ is assumed to follow a normal distribution with a mean of 0 and a variance of σ_γ^2, denoted as $\gamma \sim N(0, \sigma_\gamma^2)$, and is independent of θ. The magnitude of the passage effect is reflected by the size of σ_γ^2: the larger the σ_γ^2, the higher the degree of local dependence among the items in the passage; thus, the larger the passage effect. If σ_γ^2=0, this model reduces to the 3PL IRT model, which is used to simulate responses for the discrete items in this study.

Factors Studied in the Simulation

The following factors were manipulated in the simulation.

1. Anchor type. For the external anchor tests, different proportions of passage-based items versus discrete items were constructed at five different levels: 0% passage-based items vs. 100% discrete items, 25% vs. 75%, 50% vs. 50%, 75% vs. 25%, and 100% vs. 0%.

2. Passage effect. The variances of the testlet parameter (σ_γ^2) for passage-based items in tests *X*, *Y,* and anchor *V* were set to be equal. Two values were used: 0.2 and 0.8. While the variance of the ability θ is fixed at 1, $\sigma_\gamma^2 = 0.2$ and $\sigma_\gamma^2 = 0.8$ represented low and high passage effects, respectively. Recall that each *X* and *Y* always contains 40 passage-based items. Thus, when there is 0% passage-based items in the anchor, passage effect is still reflected through *X* and *Y*.

3. The difference in the mean ability (Δ) between population *P* and population Q . Δ took one of the three values: 0, 0.2, and 0.4. In population *P*, the ability θ always followed the standard normal distribution, θ in population *Q* followed *N*(0,1), *N*(0.2,1), or *N*(0.4,1).

4. Equating method. We used both chain equating (CE) and post-stratification equating (PSE) methods to ensure that our results are not particular to the equating method employed. CE uses the anchor as part of a chain: first link *X* to *V* in population *P*, and then link *V* to *Y* in population *Q*. The two linking functions are then composed to map *X* to *Y* through *V*. PSE uses the anchor test *V* to estimate the distribution of *X* in population *Q* and the distribution of *Y* in population *P*. It assumes that the conditional distribution of *X* given *V* and the conditional distribution of *Y* given *V* are population-invariant and then post-stratifies the distributions of both *X* and *Y* on a target population *T* (synthetic population of *P* and *Q*). Equal weights were used to form the synthetic population. For more detail on the CE and PSE methods, see Kolen and Brennan (2004) and von Davier, Holland, and Thayer (2004).

Simulation Steps

1. Generate item parameters. To make the simulated data as similar as possible to real data, item parameters were generated from distributions obtained from previous analyses of a large-scale testing program (Wang, Bradlow, & Wainer, 2002). Specifically, for test *X* and anchor *V*, we drew a sample of size 80 and another sample of size 40 from the distributions: $a_j \sim N\,(1.5,\, 0.45^2)$ while left-truncated at 0.3, $b_j \sim N\,(0,\, 1)$, and $c_j \sim N\,(0.14,\, 0.05^2)$ while left-truncated at 0.0 and right-truncated at 0.6. For test *Y*, its *a*'s and *c*'s were the same as in test *X*, while its *b*'s were 0.2 larger than those in test *X.* These three sets of item parameters were fixed across different simulation conditions and each replication in the same condition.

2. Generate ability parameters. For each population *P* and *Q,* a random sample of 4,000 θ values were drawn from ability distributions $g_P(\theta) = N(0,1)$ and $g_Q(\theta) = N(\Delta,1)$, respectively.

3. Generate testlet parameters. For each passage in *X*, *Y*, and *V*, a sample of 4,000 γ values were randomly drawn from the distribution $N(0, \sigma_{\gamma}^{2})$. For the same examinee, the same γ value holds for each item in the same passage.
4. Simulate scores on *X* in *P*, *Y* in *Q*, and *V* in both *P* and *Q*. Responses to discrete items were generated based on the item parameters and ability parameters using the 3PL IRT. For passage-based items, responses were generated based on the item parameters, ability parameters, and testlet parameters using the 3PL testlet model.
5. Presmooth each data set using a loglinear model that preserved the first six univariate moments and one cross-product moment of the bivariate distribution of total test and anchor (for detailed discussions on loglinear models, see Holland & Thayer, 1987, 2000).
6. Perform two equatings (CE and PSE) on the presmoothed data.

For each simulation condition, Steps 2 to 6 were replicated *M* = 400 times. The data generation procedure (Steps 1 to 4) was programmed using R (R Development Core Team, 2006), while presmoothing and CE and PSE equating (Steps 5 and 6) were conducted using SAS macros.

Evaluation Criteria

We used equating bias, standard error of equating (SEE) and root mean squared error (RMSE) as indices to evaluate the performance of the equatings. For each combination of σ_{γ}^{2} and Δ, samples of size 500,000 were simulated in *P* and *Q*, respectively. The two samples were combined to form a synthetic group, and a single group equipercentile equating function was calculated and was treated as the criterion population equating function. The probability of scoring a particular score *x* on form *X* in the combined sample, which is denoted as $r(x)$, was treated as the corresponding population value. Figure 1 contains $r(x)$ for different simulation conditions. The two plots on the first row are for $\sigma_{\gamma}^{2} = 0.2$ and $\sigma_{\gamma}^{2} = 0.8$, respectively, while the remaining three plots are for $\Delta = 0, 0.2,$ and 0.4, respectively.

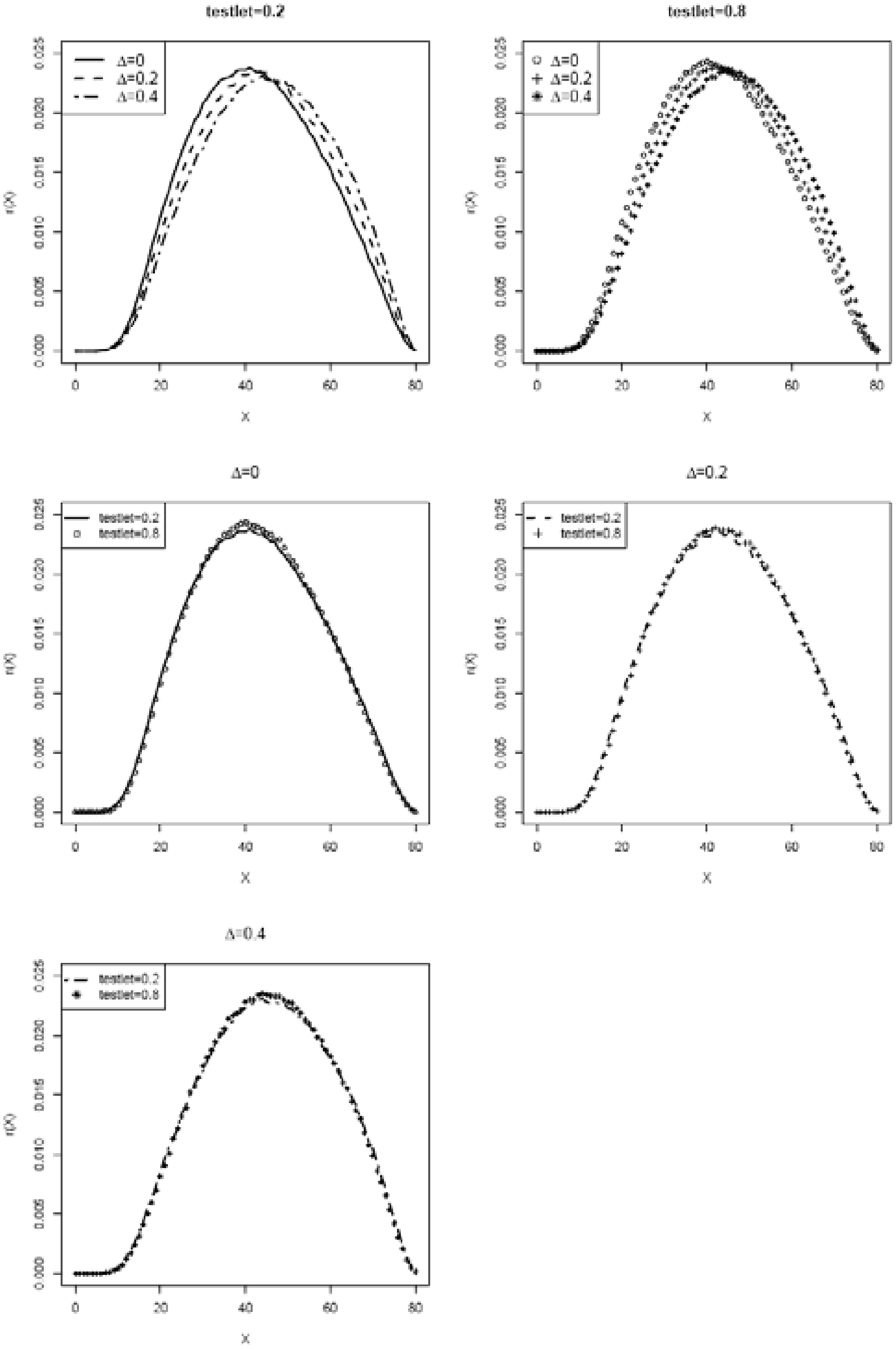


Figure 1. **Empirical distributions of X under different conditions.**

Equating bias. Equating bias is an index of systematic error of equating. Define x as a particular score on form *X*, $e(x)$ as the criterion population equating function, $\hat{e}_i(x)$ as the sample equating function that transforms scores of form *X* to the raw score scale of form *Y* in the *ith* Monte Carlo replication, and $\bar{e}(x)$ as the average of $\hat{e}_i(x)$ over the *M* number of Monte Carlo replications. Hence, the conditional bias at each score point x is calculated by

$$Bias(x) = \frac{1}{M}\sum_{i=1}^{M}[\hat{e}_i(x) - e(x)] = \bar{e}(x) - e(x), where\ \bar{e}(x) = \frac{1}{M}\sum_{i=1}^{M}\hat{e}_i(x) \quad . \tag{4}$$

As an overall measure of systematic errors across the score range, the weighted average of absolute bias is calculated by taking into account the density of scoring x in the target population: $\sum_x r(x)\,|\,Bias(x)\,|$. The absolute value of conditional bias is used for aggregation, instead of the conditional bias itself, to insure that positive and negative biases at different score levels do not cancel out.

Standard error of equating. SEE measures random error in equating that is due to sampling variability. The conditional SEE at score point x is calculated as

$$SEE(x) = \sqrt{\frac{1}{M}\sum_{i=1}^{M}[\hat{e}_i(x) - \bar{e}(x)]^2} \quad . \tag{5}$$

Similarly, the weighted average of SEE is defined as $\sqrt{\sum_x r(x) SEE^2(x)}$.

Root mean squared error. Total error in equating is the combination of random error and systematic error components. To get a measure of overall equating error, the RMSE at each score x is calculated as

$$RMSE(x) = \sqrt{\frac{1}{M}\sum_{i=1}^{M}[\hat{e}_i(x) - e(x)]^2} \quad , \tag{6}$$

and the corresponding weighted average of RMSE is $\sqrt{\sum_x r(x) RMSE^2(x)}$.

Results

Effects on Equating Bias

Table 2 presents the weighted average of absolute bias (×100). There are 5 (anchor types in terms of the percentage of passage-based items in the anchor) × 2 (passage effects) × 3 (ability differences) × 2 (equating methods) = 60 conditions. Figure 2 displays the results reported in Table 2. Plots of conditional biases for two anchor type conditions, 0% and 50% passage-based anchor items, are provided in Figure 3 and Figure 4 as examples.

Table 2

Weighted Average of Absolute Bias (×100) Under Different Conditions

Equating method	Ability difference	Passage effect	% of passage-based items in the anchor				
			0	25	50	75	100
CE	0	0.2	3	4	3	4	3
		0.8	3	3	4	4	3
	0.2	0.2	5	5	8	9	10
		0.8	3	5	11	15	24
	0.4	0.2	10	13	16	19	22
		0.8	8	12	24	33	48
PSE	0	0.2	3	3	3	4	4
		0.8	3	3	3	3	3
	0.2	0.2	30	32	37	39	42
		0.8	29	39	50	58	71
	0.4	0.2	63	68	74	80	86
		0.8	61	82	103	119	144

Anchor type effects. The effects of anchor types are reflected by the slope of lines in Figure 2. Holding other factors constant, when the group ability difference $\Delta = 0$, there is no anchor type effect. However, when Δ is not minimal, anchor tests containing more passage-based items tend to produce larger bias, especially when the passage effect σ_γ^2 is large and when PSE is used for equating.

Passage effects (σ_γ^2). Passage effects can be seen by comparing the solid line (σ_γ^2 = 0.2) and the dashed line (σ_γ^2 = 0.8) in each plot of Figure 2. There is no impact of passage on bias when the group ability difference $\Delta = 0$. However, when Δ increases, a larger passage effect produces larger bias.

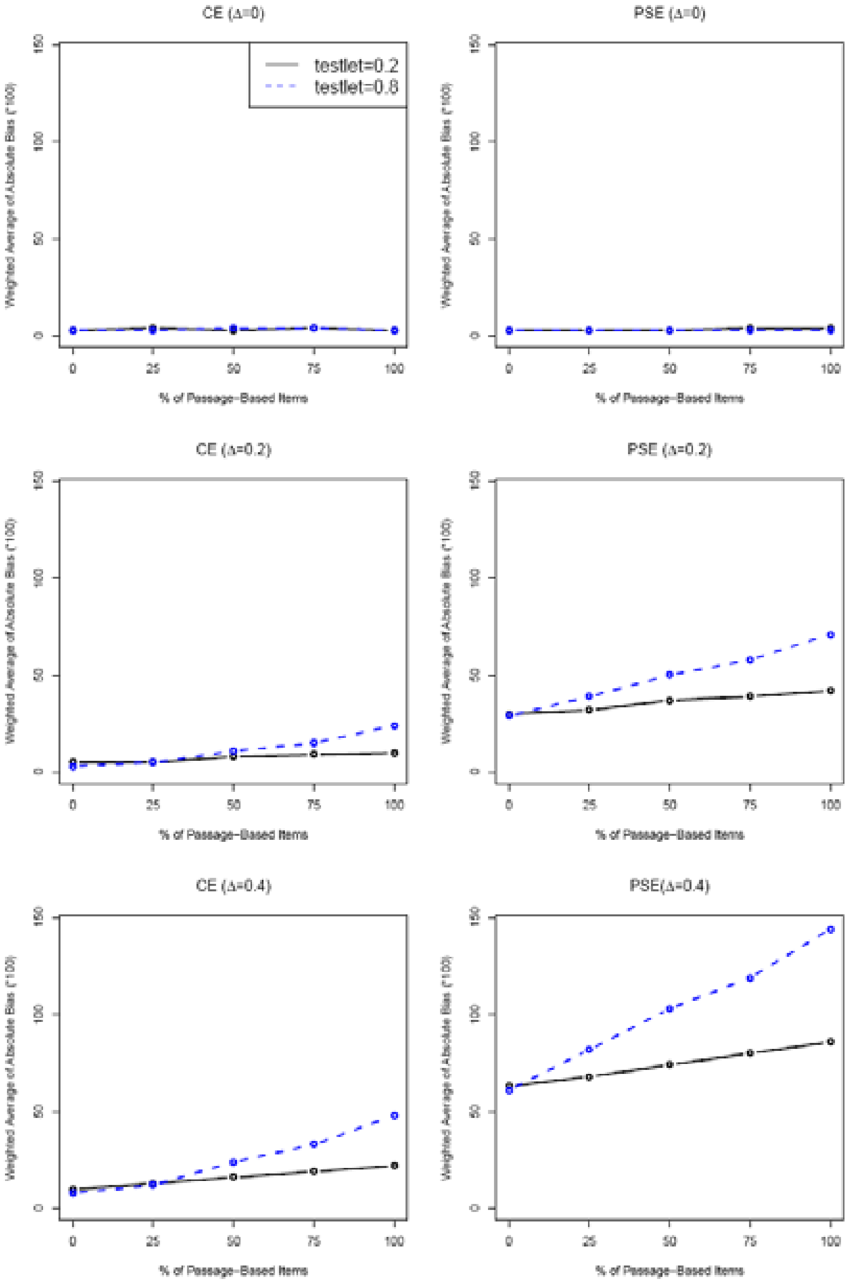
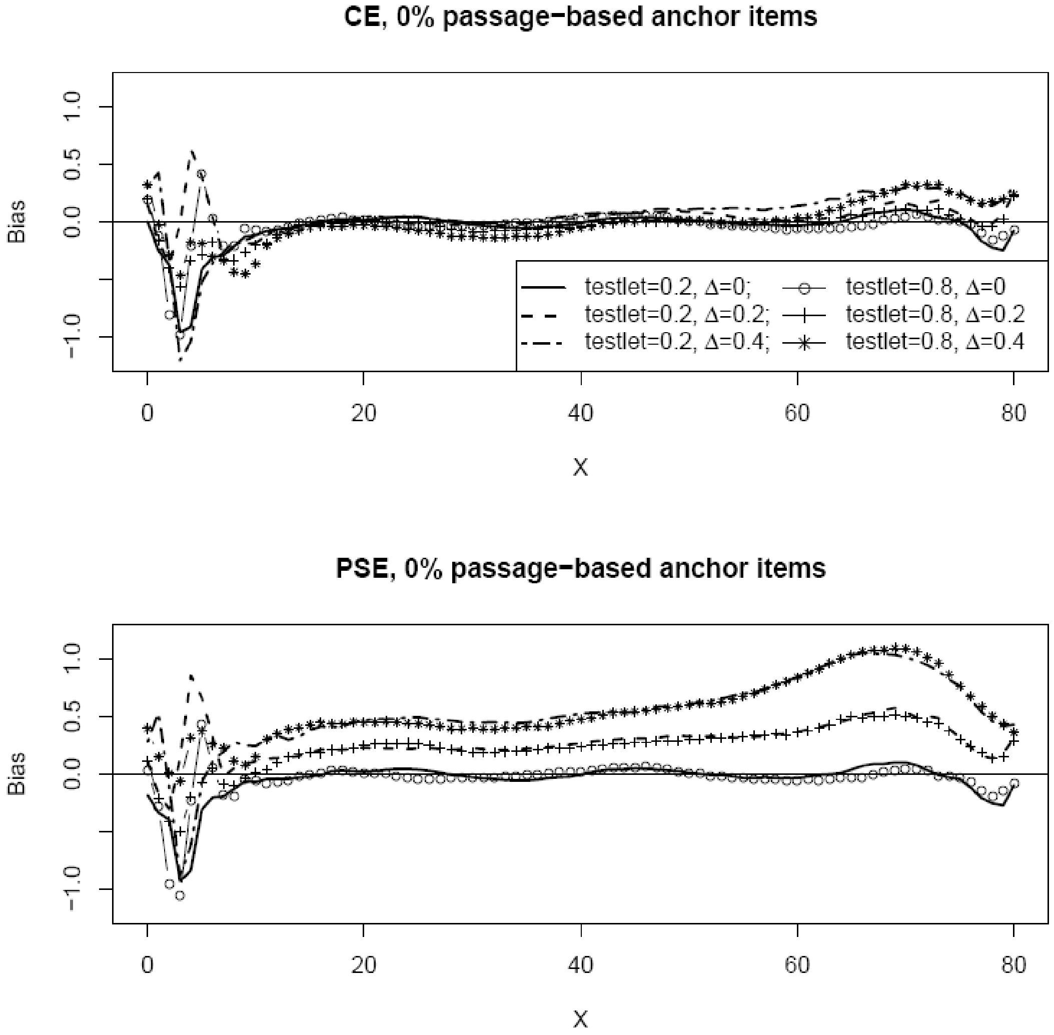


Figure 2. **Weighted average of absolute bias (×100) under different conditions.**


CE, 0% passage−based anchor items

PSE, 0% passage−based anchor items

Figure 3. **Conditional bias for 0% passage-based anchor items.**

Group ability difference (Δ) effects. Group ability difference effects on bias are reflected by comparing plots in the same column but different rows in Figure 2. Results suggested that when the group ability difference Δ=0, equating is essentially unbiased. When Δ increases, the bias increases as well. When Δ=0.4, for example, the size of the bias is nearly three times larger than that of Δ=0 for the CE method, and almost 20 times larger for the PSE method. This finding is consistent with previous research showing that group differences lead to equating bias (Kolen & Brennan, 2004; Sinharay & Holland, 2007; Wang, Lee, Brennan, & Kolen, 2008).

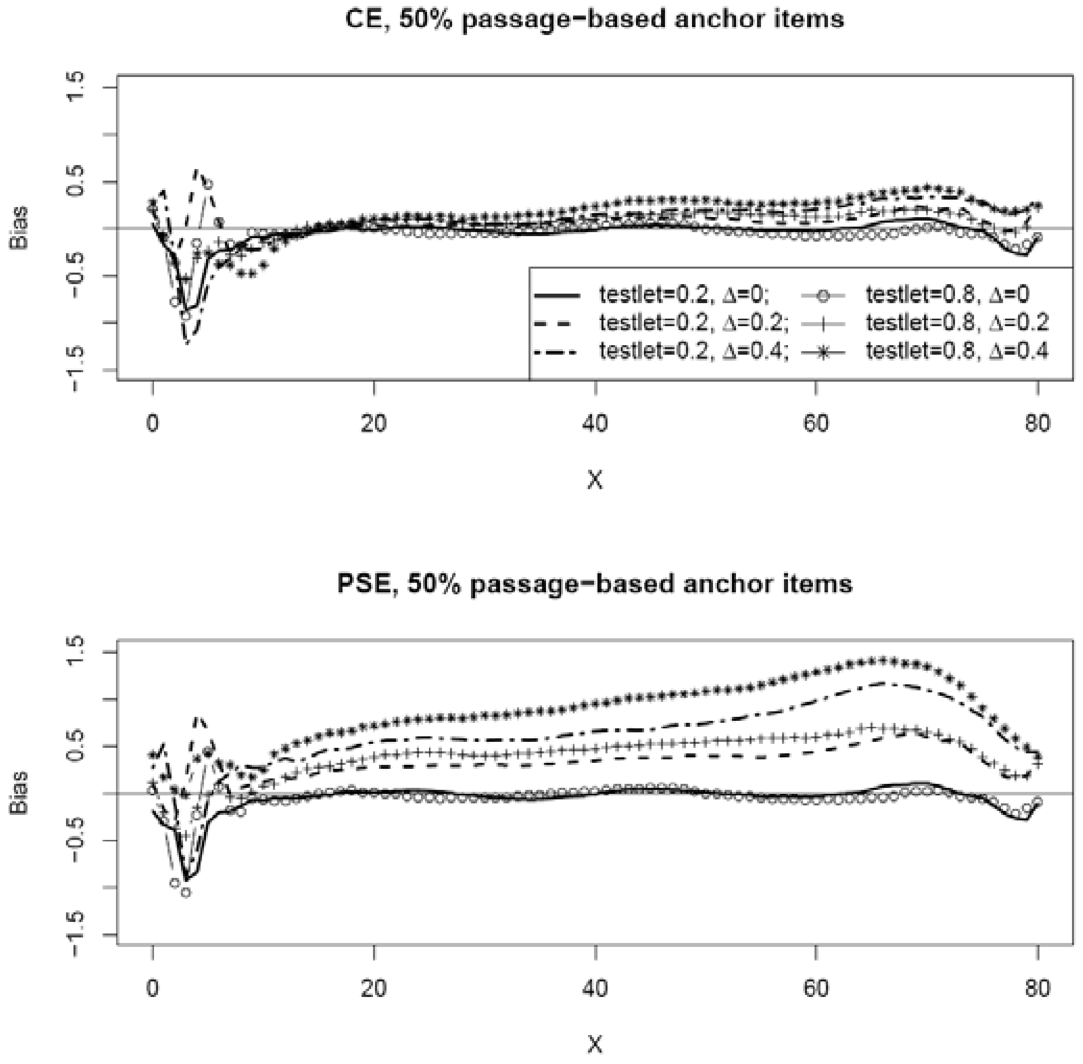


Figure 4. **Conditional bias for 50% passage-based anchor items.**

Equating method effects. By comparing graphs in the same row but different columns in Figure 2, we can see that equating bias from the CE method is less than or equal to equating bias from the PSE. This difference of bias between the two equating methods becomes larger when the group difference gets larger.

The interaction. The ability difference interacts with passage effects, with anchor types, and with equating methods. When the ability difference $\Delta = 0$, the size of the passage effect, the equating method, and the proportion of passage-based items in the anchor do not seem to matter. When the Δ gets larger, however, the equating bias increases when the passage effect and/or the percentage of passage-based items increases.

Effects on Standard Error of Equating

Table 3 and Figure 5 present the weighted average of the standard error of equating. Plots of conditional SEEs for two anchor type conditions are given in Figure 6 and Figure 7.

Table 3

Weighted Average of Standard Error of Equating (×100) Under Different Conditions

Equating method	Ability difference	Passage effect	% of passage-based items in the anchor				
			0	25	50	75	100
CE	0	0.2	22	22	24	24	24
		0.8	30	32	34	34	37
	0.2	0.2	28	29	30	30	30
		0.8	30	32	33	34	36
	0.4	0.2	29	30	31	32	31
		0.8	31	33	34	36	37
PSE	0	0.2	15	15	16	16	16
		0.8	27	29	30	30	32
	0.2	0.2	26	27	27	27	28
		0.8	27	28	30	31	32
	0.4	0.2	26	27	27	28	28
		0.8	28	29	30	31	32

Anchor type effects. The percentage of passage-based items in the anchor has a small but consistent impact on the SEE. In general, the more passage-based items in the anchor, the larger the SEE, especially when the passage effect (σ_γ^2) is bigger.

Passage effects. The impact of passages on the magnitude of SEE is very clear: the larger the size of the passage effect, the greater the SEE.

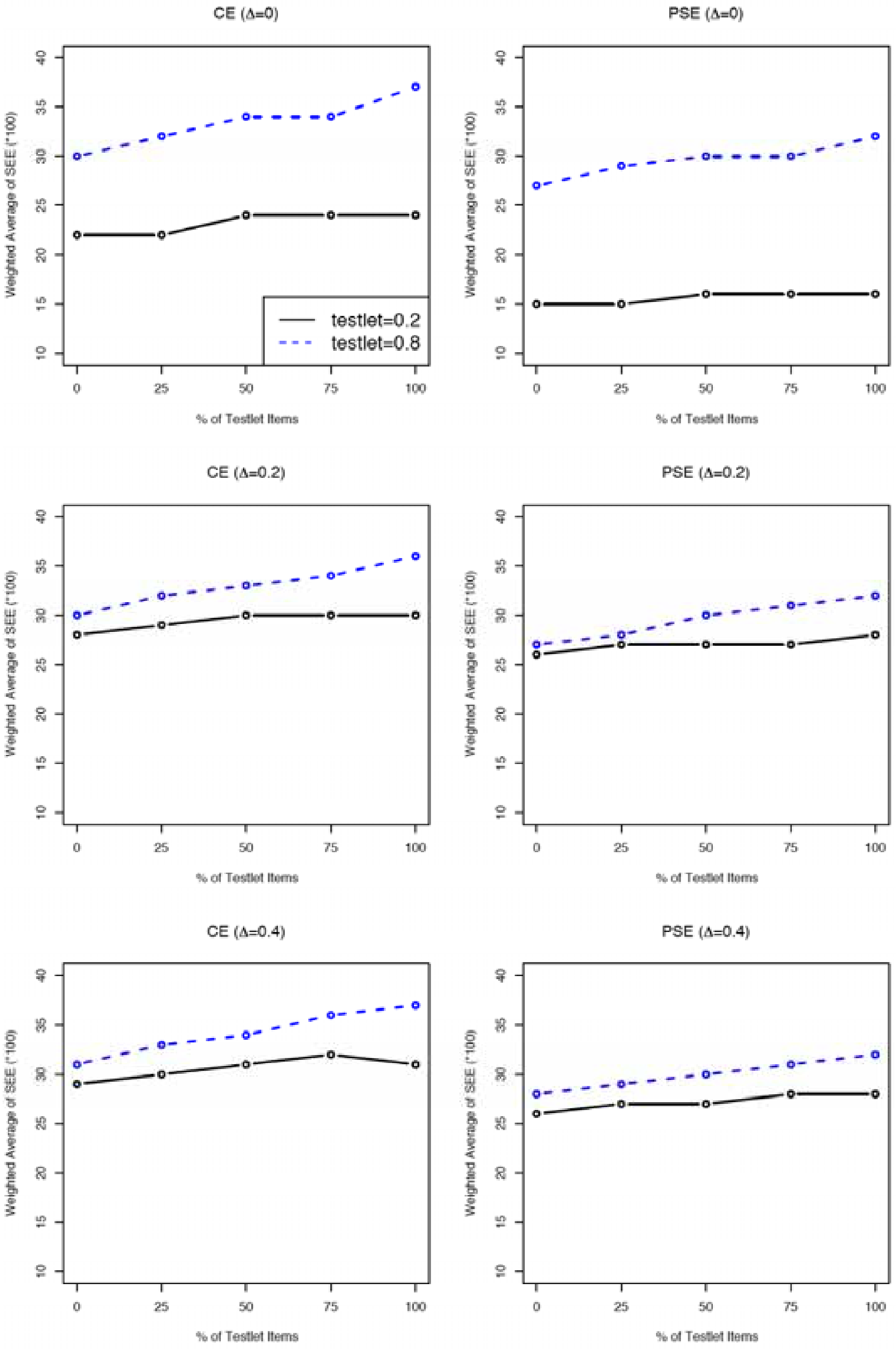
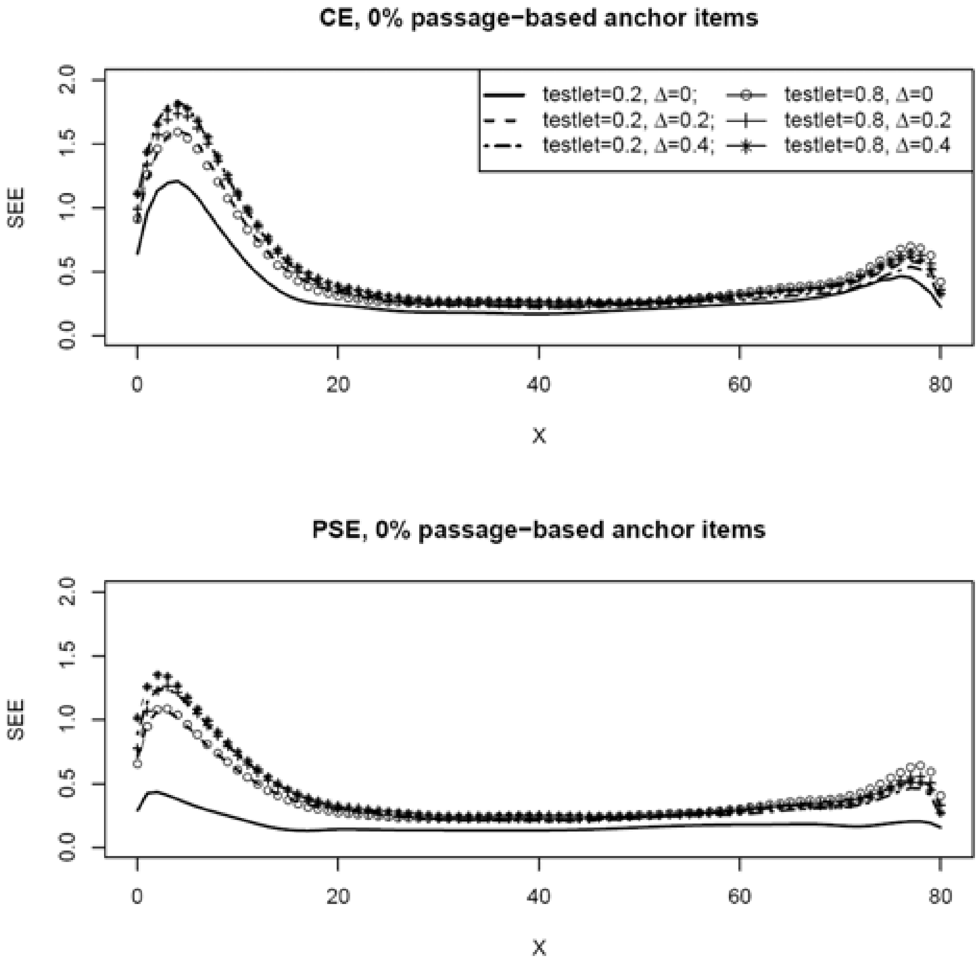


Figure 5. **Weighted average of standard error of equating (×100) under different conditions.**



CE, 0% passage−based anchor items

PSE, 0% passage−based anchor items

Figure 6. **Conditional standard error of equating for 0% passage-based anchor items.**

Group ability differences (Δ) effects. Δ has very small effects on SEE. The only exception is when Δ=0 and σ_{γ}^{2} = 0.2, SEE values appear to be relatively smaller than SEE values in other Δ conditions. Plots of conditional SEE in Figure 6 and Figure 7 further depict that the SEE at each score level, when Δ=0 and σ_{γ}^{2} = 0.2, is smaller than that of other conditions.

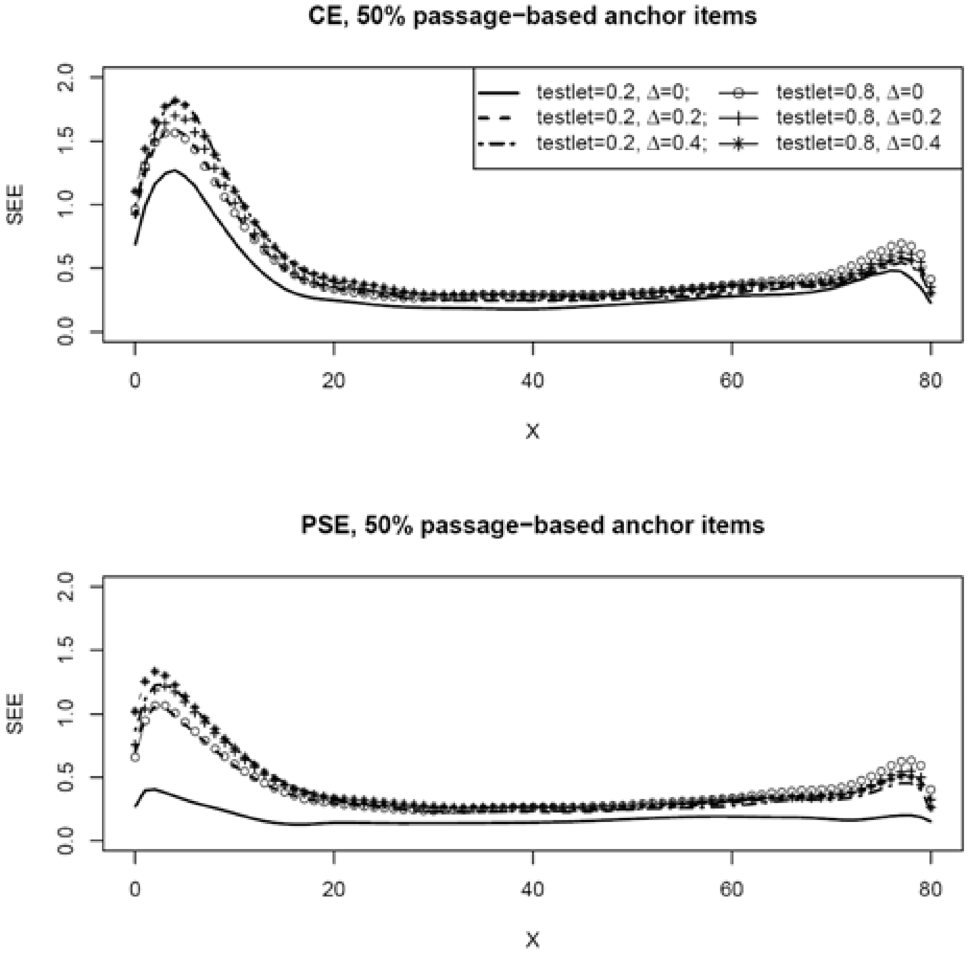


Figure 7. **Conditional standard error of equating for 50% passage-based anchor items.**

The interaction. The passage effect interacts withΔ. When Δ=0, passage effect has larger influence on SEE than when $\Delta \neq 0$.

Equating method effects. PSE consistently produces smaller SEEs than CE does.

Effects on Root Mean Squared Error

The RMSE results are presented in Table 4 and Figure 8. Plots of conditional RMSEs for two anchor type conditions are provided in Figure 9 and Figure 10 as examples.

Table 4

Weighted Average of Root Mean Squared Error (×100) Under Different Conditions

Equating method	Ability difference	Passage effect	% of passage-based items in the anchor				
			0	25	50	75	100
CE	0	0.2	22	23	24	25	25
		0.8	30	32	34	34	37
	0.2	0.2	29	30	31	32	32
		0.8	30	32	35	38	44
	0.4	0.2	31	34	35	37	39
		0.8	33	36	42	49	62
PSE	0	0.2	16	16	16	17	17
		0.8	27	29	30	30	32
	0.2	0.2	41	43	47	49	51
		0.8	40	49	59	66	79
	0.4	0.2	70	76	82	86	93
		0.8	70	89	109	125	149

Anchor type effects. When Δ=0, the anchor type virtually has no effect on RMSE. When the two populations have different average ability, RMSE increases as the proportion of passage-based items increases.

Passage effects. RMSE increases as the size of the passage effect increases.

Mean ability differences effects. Δ has an effect on RMSE whereby the larger the Δ, the larger the RMSE.

Equating method effects. PSE produces slightly better results than CE when Δ=0, whereas CE produces better results than PSE when $\Delta \neq 0$. The difference between CE and PSE increases when Δ increases.

The interaction. The ability difference Δ interacts with anchor type and with equating method. When $\Delta = 0$, the percentage of passage-based anchor items has little effect on RMSE, and PSE performs better than CE. However, when the Δ gets larger and the percentage of passage-based anchor items increases, the RMSE also increases; moreover, CE performs better than PSE.

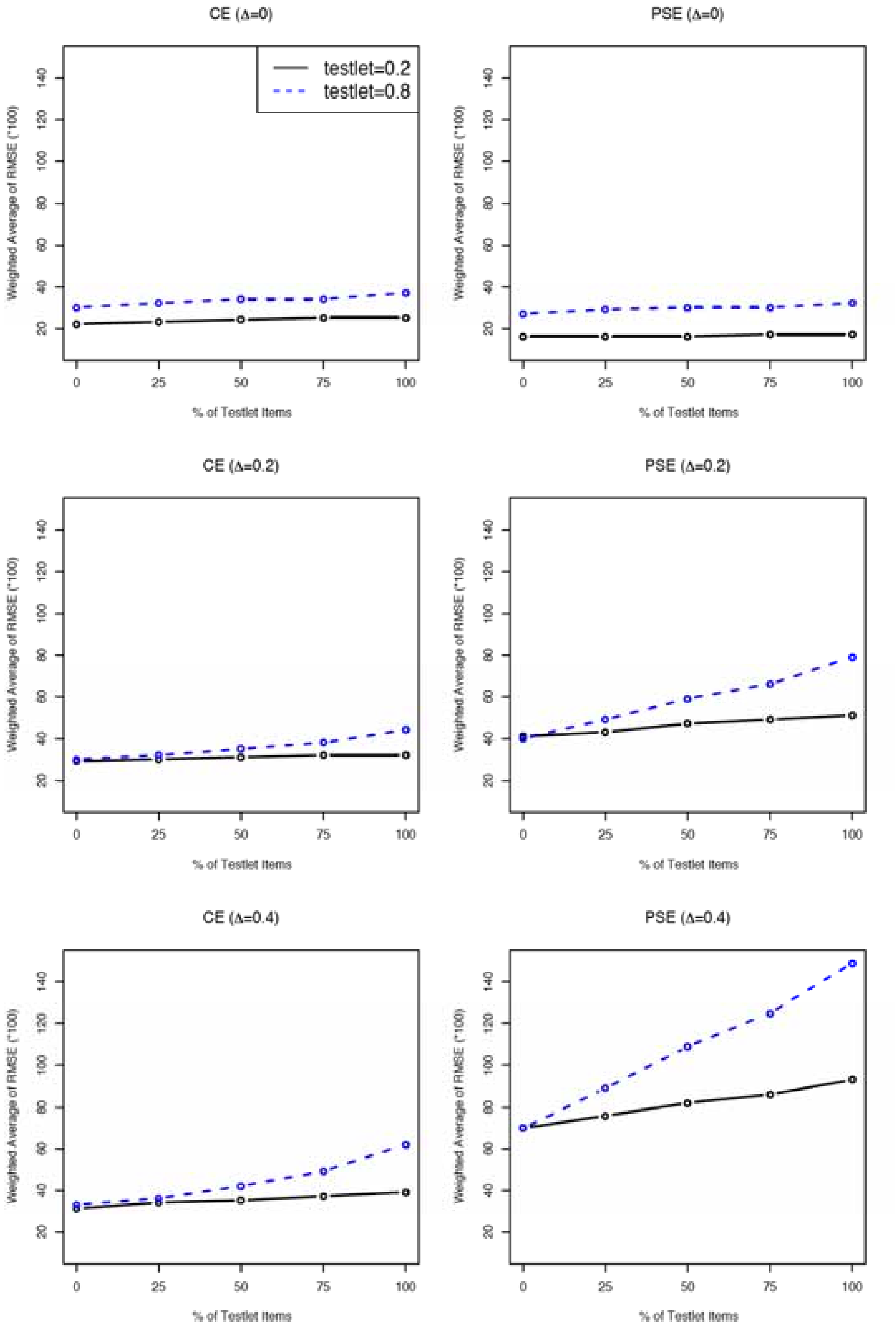


Figure 8. **Weighted average of root mean squared error (×100) under different conditions.**

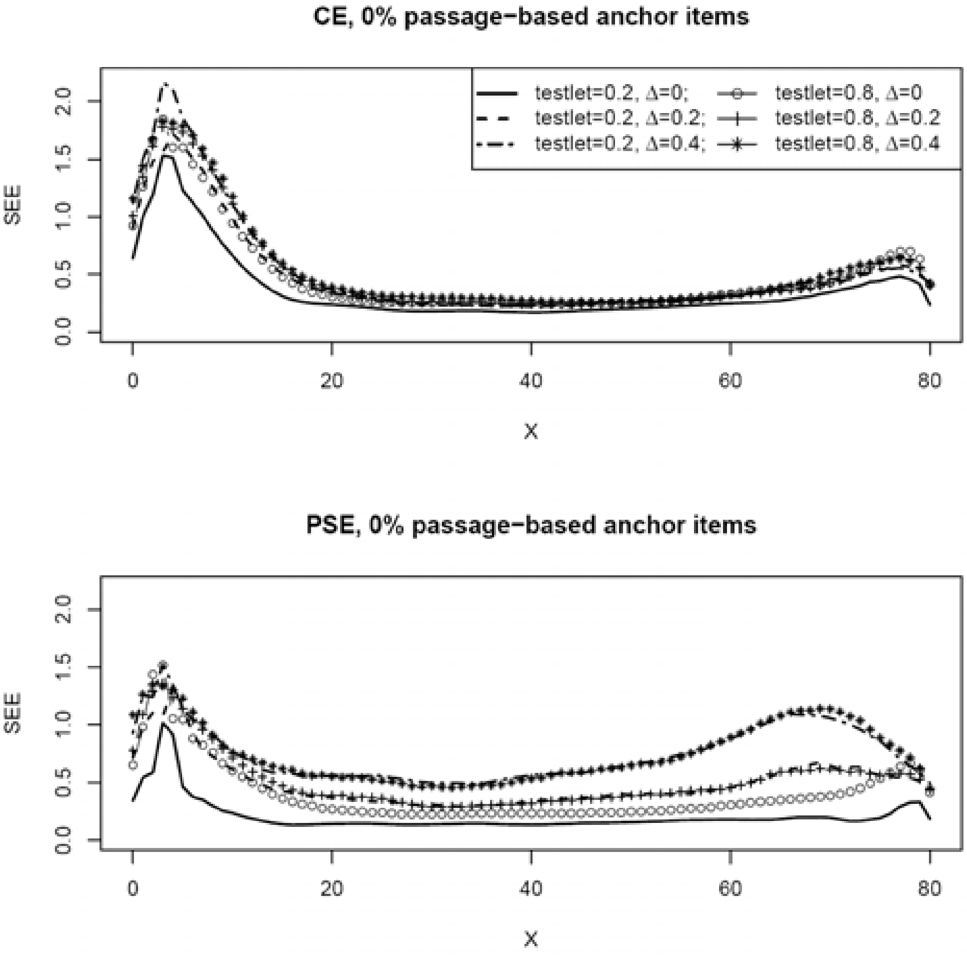


Figure 9. **Conditional root mean squared error for 0% passage-based anchor items.**

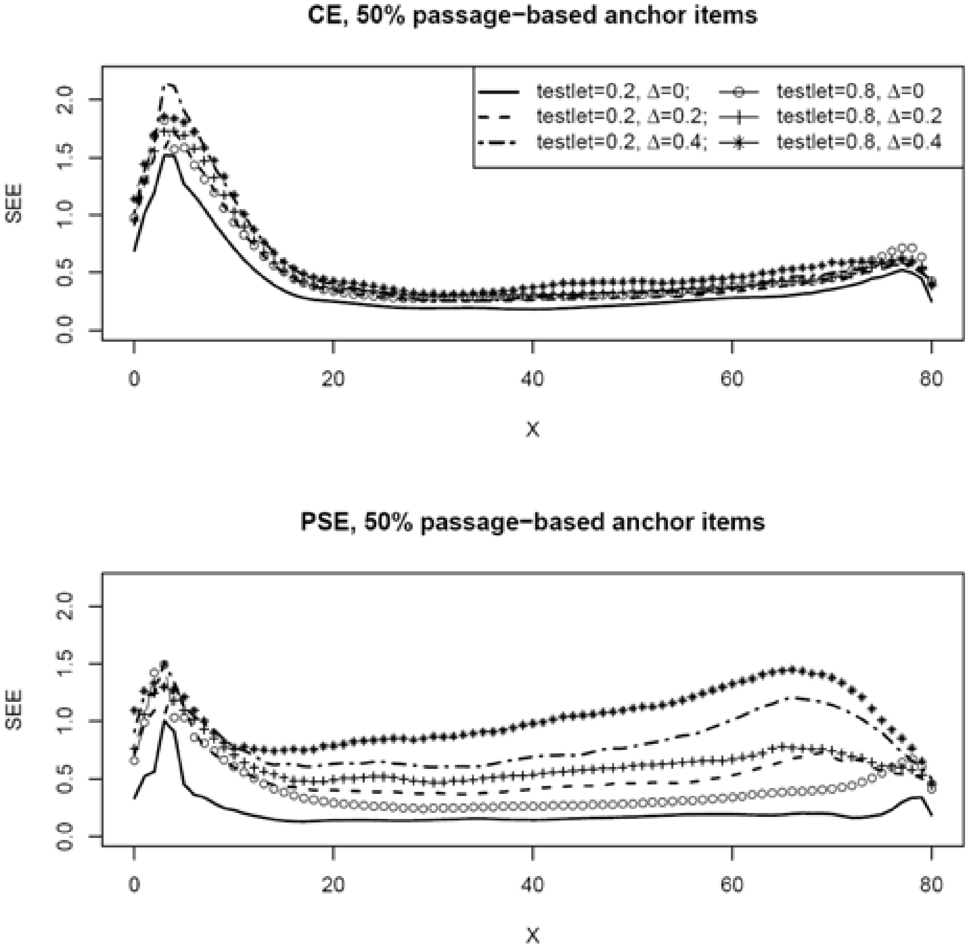


Figure 10. **Conditional root mean squared error for 50% passage-based anchor items.**

Conclusions and Discussion

This study explores the impact of two different types of anchor item on observed score equating. These two types are: discrete items, each of which is based on a unique item stem, and passage-based items that share one common stem. Using simulated data, we compared equating performance with respect to systematic equating error (bias), random equating error (standard error of equating), and total equating error (RMSE). In addition to studying various combinations of discrete and passage-based items in anchor tests and differing degrees of passage effects, we also studied two other factors: ability difference between the two populations and type of equating method.

Our main conclusions are that when the group ability difference is zero, anchor types, passage effects, and equating methods have little effect on equating bias, but standard error of equating is smaller when the proportion of passage-based items or passage effect is smaller. When the group ability difference is not zero, an anchor with more discrete items and/or with smaller passage effects produces less systematic and random equating errors. Regarding equating methods, CE performs better than PSE under these conditions.

Our findings suggest that an anchor with more discrete items would be preferred, even if such an anchor may not be a miniature version of the total test. (Recall that the total test has 50% discrete items and 50% passage-based items). There are two potential reasons for the less than optimal equating performance when employing passage-based anchor items. First, because passage-based items tend to be locally dependent, each item provides less unique information about the latent ability. Second, passage-based items may measure a construct other than what the test is designed to measure. For example, the nature of reading may be affected by variables such as reader's background knowledge of subject matter or cultural knowledge (Alderson, 2000). These two characteristics of passage-based items lower the reliability of the test and the correlation between the total test and the anchor, which lead to more equating errors.

Our findings will be most applicable to testing programs that struggle with the use of passage-based items in the anchor test. Passage-based items, or the reading passages, are much easier to memorize than a series of discrete items. Once the passage-based anchor items get compromised, the actual equating results will be consequently contaminated. Yet many testing programs still use passage-based items as anchor items because it is widely believed that the anchor should be a miniature version of the total test: if the total test contains passage-based items, so should the anchor. Our study provides evidence to practitioners that an anchor containing a proportion of passage-based items the same as the proportion of passage-based items in the total test may not necessarily produce the best equating results.

Our study is an initial attempt to investigate the impact of passage-based items on equating. Future work remains to be done. First, simulations can be designed to take into account the possibility that examinees in different populations may respond differently to discrete and passage-based items. For example, some examinees may perform better—relative to the group of all examinees—on discrete items than on passage-based items, while other examinees perform better on passage-based items than on discrete items. For a simplified example, we call the two types of examinees *Type P* and *Type D* to indicate their preference for passage-based or discrete items. Suppose the new-form sample includes a substantially greater percentage of Type P examinees than the old-form sample. An anchor with no passage-based items will make the Type P examinees look weaker than they are and will make the Type D examinees look stronger than they are. The new-form sample, with more Type P examinees, will appear to be less able than it really is. As a consequence, the new test will appear to be somewhat easier than it really is relative to the old test. Therefore, the scores reported on the new test will be lower than they should be. The equating will be biased in a downward direction. The relevant question, for someone selecting anchor items and selecting equating samples, is whether it is safe to assume that this combination of conditions will not occur. That is an empirical question, and the results may not generalize beyond the specific test and conditions that were investigated.

Second, more flexible models than the 3PL testlet model can be used to represent local dependence among passage-based items. As Li, Bolt, and Fu (2006) pointed out, the 3PL testlet model assumes the latent proficiency and passage dimensions have the same discrimination parameter, which may not hold in reality. Their results of analyzing real data suggested that a model with separate discrimination parameters for different dimensions provides a better fit to the data.

Third, this project only studied the impact of passage-based items on observed score equating. Whether IRT equating methods (e.g., IRT true score and IRT observed score equating) are robust to passage effects is also worthy of investigating. Finally, equating results of using passage-based and discrete anchor items should be examined based on multiple operational data sets.

References

Alderson, J. C. (2000). *Assessing reading*. UK: Cambridge University Press.

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (ETS Research Rep. No. RR-87-31). Princeton, NJ: ETS.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25*, 133–183.

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

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

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30,* 3-21.

Liu, J., Sinharay, S., Holland, P., Feigenbaum, M., & Curley, E. (2009). *The effects of different types of anchor tests on observed score equating* (ETS Research Rep. No. RR-09-41). Princeton, NJ: ETS.

Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: ETS.

R Development Core Team. (2006). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved June, 2008, from http://www.R-project.org

Rijmen, F. (2009). *Three multidimensional models for testlet based tests: Formal relations and an empirical comparison* (ETS Research Rep. No. RR-09-37). Princeton, NJ: ETS

Sinharay, S., & Holland, P. W. (2006). *The correlation between the scores of a test and an anchor test* (ETS Research Rep. No. RR-06-04). Princeton, NJ: ETS.

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.

von Davier, A., Holland, P.W., & Thayer, D. (2004). The chain and post-stratification methods for observed-score equating: Their relationship to population invariance. *Journal of Educational Measurement, 41*(1), 15-32.

Wainer, H., Bradlow, E. T., & Wang X. (2007). *Testlet response theory and its applications*. UK: Cambridge University Press.

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

Wang, X., Bradlow, E. T., & Wainer, H. (2002). *A general Bayesian model for testlets: Theory and applications* (ETS Research Rep. No. RR-02-02). Princeton, NJ: ETS.[

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