

Is It Necessary to Make Anchor Tests Mini-Version of the Tests Being Equated or Can Some Restrictions Be Relaxed?

Sandip Sinharay and Paul W. Holland
ETS, Princeton, NJ

It is a widely held belief that anchor tests should be miniature versions (i.e., minitests), with respect to content and statistical characteristics, of the tests being equated. This article examines the foundations for this belief regarding statistical characteristics. It examines the requirement of statistical representativeness of anchor tests that are content representative. The equating performance of several types of anchor tests, including those having statistical characteristics that differ from those of the tests being equated, is examined through several simulation studies and a real data example. Anchor tests with a spread of item difficulties less than that of a total test seem to perform as well as a minitest with respect to equating bias and equating standard error. Hence, the results demonstrate that requiring an anchor test to mimic the statistical characteristics of the total test may be too restrictive and need not be optimal. As a side benefit, this article also provides a comparison of the equating performance of post-stratification equating and chain equipercentile equating.

The Non-Equivalent groups with Anchor Test (NEAT) design is one of the most flexible tools available for equating tests (e.g., Angoff, 1971; Kolen & Brennan, 2004; Petersen, Marco, & Stewart, 1982; Petersen, Kolen, & Hoover, 1989). The NEAT design deals with two non-equivalent groups of examinees and an anchor test. The design table for a NEAT design is shown in Table 1.

The test X corresponds to the new form given to a sample from population P and the test Y corresponds to the old form given to a sample from population Q . The anchor test A is given to both P and Q . The choice of anchor test is crucial to the quality of equating with the NEAT design.

It is a widely held belief that an anchor test should be a miniature version (i.e., a *minitest*) of the tests being equated. Angoff (1968, p. 12) and Budescu (1985, p. 15) recommended an anchor test that is a *parallel miniature of the operational forms*. More specifically, it is recommended that an anchor test be proportionally representative or a mirror of the total tests in both *content* and *statistical* characteristics (Dorans, Kubiak, & Melican, 1998, p. 3; Kolen and Brennan, 2004, p. 19; Petersen et al., 1989, p. 246; von Davier, Holland, & Thayer, 2004, p. 33). Currently, most operational testing programs that use the NEAT design employ a minitest as the anchor; to ensure statistical representativeness, the usual practice is to make sure that the mean and spread of the item difficulties of the anchor test are roughly equal to those of the tests being equated (see, e.g., Dorans et al., p. 5).

The requirement that the anchor test be representative of the total tests (i.e., the tests being equated) with respect to content¹ is justified from the perspective of

TABLE 1
The NEAT Design

Population	New Form <i>X</i>	Old Form <i>Y</i>	Anchor <i>A</i>
New form population <i>P</i>	✓		✓
Old form population <i>Q</i>		✓	✓

content validity and has been shown to be important by Klein and Jarjoura (1985) and Cook and Petersen (1987). Peterson, Marco, and Stewart (1982) demonstrated the importance of having the mean difficulty of the anchor tests close to that of the total tests. We also acknowledge the importance of these two aspects of an anchor test. However, the literature does not offer any proof of the superiority of an anchor test for which the spread of the item difficulties is representative of the total tests. Furthermore, a minitest has to include very difficult or very easy items to ensure adequate spread of item difficulties, which can be problematic as such items are usually scarce (one reason being that such items often have poor statistical properties like low discrimination and are thrown out of the item pool). An anchor test that relaxes the requirement on the spread of the item difficulties might be more operationally convenient, especially for testing programs using external anchor tests.

Motivated by the above, this article focuses on anchor tests that

- are content representative
- have the same mean difficulty as the total tests
- have spread of item difficulties less than that of the total tests

Operationally, such an anchor test can be constructed exactly in the same manner as the minitests are constructed except for the requirement that it mimic the spread of the item difficulties of the total tests. Because items with moderate difficulty values are often more common, an operationally convenient strategy to construct such an anchor test may be to include only moderate-difficulty items in the anchor test.

To demonstrate the adequate performance of anchor tests with spread of item difficulties less than that of the minitest, Sinharay and Holland (2006) defined a “midtest” as an anchor test with a very small spread of item difficulties and a “semi-midtest” as one with a spread of item difficulty that lies between those of the midtest and the minitest. The semi-midtests will often be easier to construct operationally than minitests because there is no need to include very difficult or very easy items in them. Sinharay and Holland cited several works that suggest that the midtest will be satisfactory with respect to psychometric properties like reliability and validity. The next step is to examine how these anchor tests perform compared to the minitests in test equating.

Sinharay and Holland (2006), using a number of simulation studies and a real data example, showed that the midtests and semi-midtests have slightly higher anchor-test-to-total-test correlations than the minitests. As higher anchor-test-to-total-test

correlations are believed to lead to better equating (Angoff, 1971, p. 577; Dorans et al., 1998; Petersen et al., 1989, p. 246; etc.), the findings of Sinharay and Holland (2006) suggest that a minitest may not be the optimum anchor test, and beg for a direct comparison of the equating-performance of minitests versus midtests and semi-midtests. Hence, the present article compares the equating performance of minitests versus that of midtests and semi-midtests through a series of simulation studies and a pseudo-data example.

The next section compares the minitests versus the other two types of anchor tests for a simple equating design. The following two sections compare the equating performance of the minitest and the other anchor tests in the context of NEAT design using data simulated from unidimensional and multidimensional item response theory (IRT) models. The penultimate section describes similar results for a pseudo-data example. The last section provides discussion and conclusions.

Comparison of Minitests and Other Anchor Tests for a Simple Equating Design

Consider the simple case of a random groups design with anchor test (Angoff, 1971; Kolen & Brennan, 2004; Lord, 1950) in which randomly equivalent groups of examinees are administered one of two tests that include an anchor test. Denote by X and Y the tests to be equated and the anchor test by A . Under the assumptions that (i) the populations taking tests X and Y are randomly equivalent, (ii) scores on X and A , and on Y and A are bivariate normally distributed, (iii) the correlation between scores in X and A is equal to that between scores in Y and A , and (iv) sample sizes for examinees taking the old and new forms are equal, Lord (1950) shows that the square of the standard error of equating (SEE) at any value x_i of score in test X can be approximated as

$$\text{Var}(\hat{I}_Y(x_i)) \approx \frac{\sigma_Y^2}{N} \left[2(1 - \rho_{XA}^2) + (1 - \rho_{XA}^4) \left(\frac{x_i - \mu_X}{\sigma_X} \right)^2 \right], \quad (1)$$

where the symbols have their usual meanings. Equation 1 shows that as the anchor-test-to-total-test correlation ρ_{XA} increases, the SEE decreases (a phenomenon mentioned by Budescu, 1985, p. 15). Therefore, higher correlations between X and A will result in lower SEEs in this case. This basic fact emphasizes the importance of the results of Sinharay and Holland (2006) that focuses only on ρ_{XA} as a surrogate for the more detailed study of equating in the present article.

Sinharay and Holland (2006) considered a basic skills test with 110 multiple choice items that was administered to 6,489 examinees. They used the operational 34-item internal anchor test as a minitest. A 34-item semi-midtest was formed. For the test data set, $N = 6489$, $\widehat{\mu}_X = 77.5$, $\widehat{\sigma}_X = 10.8$, and the values of $\widehat{\rho}_{XA}$ were .875 and .893, respectively, for the minitest and the semi-midtest. The SEEs for the semi-midtest and minitest were computed using the above-mentioned values and under the additional assumptions that $\sigma_Y = \sigma_X$ and $\rho_{XA} = \rho_{YA}$. The SEE for the semi-midtest is always less than that for the minitest. The percent reduction in SEE for the semi-midtest ranges from 6% to 7%.

No result as simple as Equation 1 is found for the other popular equating designs, especially for the NEAT design. Furthermore, in addition to sampling variability (or SEE), it is also important to examine the other major type of equating error: the systematic error or *equating bias* (Kolen & Brennan, 2004, p. 231). There are no general results for measuring equating bias. Hence, the next section reports the results of a detailed simulation study under a NEAT design that was performed to investigate both equating bias and variability under several conditions.

Simulations Under a NEAT Design from a Unidimensional IRT Model

Here, we simulated data from the two-parameter logistic (2PL) model, with the item response function (IRF)

$$\frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}, \tag{1}$$

where the symbols have their usual meanings.

Simulation Design

Factors controlled in the simulation. We varied the following factors in the simulations:

1. “Test length”. X and Y are always of equal length that is one of the values 45, 60, and 78, to emulate three operational tests: (i) a 45-item basic skills test, (ii) the 60-item mathematics section of an admissions test, and (iii) the 78-item verbal section of the same admissions test. The factor that we henceforth denote by test length refers to more than simply the length of the tests to be equated. Each test length has its own set of item parameters that was created to emulate those of the operational test data set on which it is based (see below). Moreover, the length of the anchor test for each test length is different as indicated in point 5, below. For this reason we put quotes around “test length.”
2. Sample size. The sample sizes for P and Q are equal and are equal to one of three values: 100 (small), 500 (medium), and 5,000 (large).
3. The difference in the mean ability (denoted as Δ_a) of the two examinee populations P and Q . Four values were used: $-.2$, 0 , $.2$, and $.4$. Units are in standard deviation (SD) of θ .
4. The difference in the mean difficulty (denoted as Δ_d) of the two tests X and Y . Three values were used: 0 , $.2$, and $.5$. Units are in SD of θ .
5. The anchor test. We constructed a minitest, a semi-miditest, and a miditest, by varying the SD of the generating difficult parameters. The SD of the difficulty parameters of the minitest, the semi-miditest, and the miditest were assumed to be, respectively, 100%, 50%, and 10% of the SD of the difficulty of the total tests. The anchor test is of length 20 for the 45-item basic skills test, and is of the same length as in the operational administrations of the two admissions tests—35 for the 78-item test and 25 for the 60-item test.

The average difficulty of the anchor tests was always centered at the average difficulty level of Y , the old test form. We did not set the average difficulty of the anchor tests at the average of the difficulty levels of X and Y . In operational testing, the usual target is to make X of the same difficulty as Y . The test X often ends up being easier or more difficult than Y because of unforeseen reasons and one can rarely anticipate the difficulty of X beforehand.

6. The equating method. To make sure that our conclusions do not depend on the equating method, we used two equipercentile equating methods, the post-stratification equating (PSE) and chain equating (CE) methods. While applying the PSE method, the synthetic population was formed by placing equal weights on P and Q .

The values for the above six factors were chosen after examining data from several operational tests.

Generating item parameters for the total and anchor tests. The 2PL model was fitted to a data set from each of the above-mentioned three operational tests to obtain marginal maximum likelihood estimates of the item parameters under a $\mathcal{N}(0, 1)$ ability distribution. These three sets of item parameter estimates were used as generating item parameters of the three Y -tests. Then, a bivariate normal distribution \mathcal{D} was fitted to the estimated $\log(a)$'s and b 's for each Y -test. The generating item parameters for each X -test were drawn from the respective fitted bivariate normal distribution \mathcal{D} . Then the difficulty parameters for X were all increased by the amount Δ_d to ensure that the difference in mean difficulty of X and Y is equal to Δ_d . The generating item parameters for the anchor tests were also drawn using the fitted distribution \mathcal{D} . The generating item parameters of the minitest were drawn from the distribution \mathcal{D} as is. The item parameters of the semi-miditest and the miditest were generated from a distribution that is the same as \mathcal{D} except for the SD of the difficulties, which was set to one-half and one-tenth, respectively, of the corresponding quantity in \mathcal{D} . Note that the generating a -parameters are obtained by taking an exponential transformation of the generated $\log(a)$'s. The generating item parameters for the X -test, Y -test, and the anchor tests were the same for all M replications under a simulation condition. For any combination of a "test length," Δ_a , and Δ_d , the *population equipercentile equating function* (PEEF), described shortly, is computed and used as the criterion.

The steps in the simulation. For each simulation condition (determined by a "test length," sample size, Δ_a , and Δ_d), the generating item parameters of the Y -test were the same as the estimated item parameters from the corresponding real data, and the generating item parameters of the X -test and the anchor tests were randomly drawn (as described earlier) once, and then $M = 1,000$ replications were performed. Each replication involved the following three steps:

1. Generate the ability parameters θ for the populations P and Q from ability distributions $g_P(\theta) = \mathcal{N}(\Delta_a, 1)$ and $g_Q(\theta) = \mathcal{N}(0, 1)$, respectively.
2. Simulate scores on X in P , Y in Q , and those on the minitest, miditest, and semi-miditests for both P and Q from a 2PL model using the draws of θ from step 1, the fixed item parameters for Y , and the generated item parameters for X and the anchor tests.

3. Perform six equatings using the scores of X in P , Y in Q , and those of the minitest, midtest, and semi-midtest in P and Q . One equating is done for each combination of an anchor test (of the three) and an equating method (either PSE or CE). Each of these equatings involved (i) presmoothing the observed test-anchor test bivariate raw score distribution using a loglinear model (Holland & Thayer, 2000) that preserved the first five univariate moments and a crossproduct moment (increasing the number of moments did not affect the results substantially), and (ii) equipercentile equating with linear interpolation (e.g., Kolen & Brennan, 2004) to continuize the discrete score distributions.

Computation of the population equipercentile equating function. The PEEF for any combination of a “test length”, Δ_a , and Δ_d was the single-group equipercentile equating of X to Y using the *true* raw score distribution of X and Y in a synthetic population T that places equal weights on P and Q . We used the iterative approach of Lord and Wingersky (1984) to obtain $P(X = x|\theta)$, the probability of obtaining a raw score of $X = x$ by an examinee with ability θ . This required the values of the item parameters and we used the generating item parameters for test X . Once $P(X = x|\theta)$ is computed, $r(x)$, the probability of a raw score of x on test X in population T is obtained by numerical integration as

$$r(x) = \int_{\theta} P(X = x|\theta)g_T(\theta)d\theta, \tag{2}$$

where $g_T(\theta) = .5g_P(\theta) + .5g_Q(\theta)$. The same approach provided us with $s(y)$, the probability of a raw score of y on test Y in population T . The true raw score distributions $r(x)$ and $s(y)$, both discrete distributions, are then continuized using linear interpolation (e.g., Kolen & Brennan, 2004). Let us denote the corresponding continuized cumulative distributions as $R(x)$ and $S(y)$, respectively. The PEEF is then obtained as $S^{-1}(R(x))$. The PEEF is the same for each replication and sample size, but varies with “test length,” Δ_a , and Δ_d . The PEEF can be seen as the population value of the IRT observed score equating (e.g., Kolen & Brennan, 2004) using linear interpolation as the continuization method.

Computation of the performance criteria: equating bias, SD, and RMSE. After the equating results from the M replications are obtained, we compare the anchor tests using bias (a measure of systematic error in equating) and SD (a measure of random error in equating) as performance criteria. For a simulation condition, let $\hat{e}_i(x)$ be the equating function in the i th replication providing the transformation of a raw score point x in X to the raw score scale of Y . Suppose $e(x)$ denotes the corresponding PEEF. The bias at score-point x is obtained as

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

and the corresponding standard deviation is obtained as

$$SD(x) = \left\{ \frac{1}{M} \sum_{i=1}^M [\hat{e}_i(x) - \bar{e}(x)]^2 \right\}^{\frac{1}{2}}.$$

We can also compute the corresponding root mean squared error (RMSE) as

$$RMSE(x) = \left\{ \frac{1}{M} \sum_{i=1}^M [\hat{e}_i(x) - e(x)]^2 \right\}^{\frac{1}{2}}.$$

It can be shown that

$$[RMSE(x)]^2 = [SD(x)]^2 + [Bias(x)]^2,$$

i.e., the RMSE combines information from the random and systematic error.

As overall summary measures for each simulation case, we compute the weighted average of bias, $\sum_x r(x)Bias(x)$, the weighted average of SD , $\sqrt{\sum_x r(x)SD^2(x)}$, and the weighted average of RMSE, $\sqrt{\sum_x r(x)RMSE^2(x)}$, where $r(x)$ is defined in Equation 2.

How realistic are our simulations? To have wide implications, it is important that our simulations produce test data that adequately reflect reality. Hence, we used real data as much as possible in our simulations from a unidimensional IRT model. Further, Davey, Nering, and Thompson (1997, p. 7) reported that simulation under an unidimensional IRT model reproduces the raw score distribution of real item response data quite adequately. The data sets simulated in our study were found to adequately reproduce the raw score distributions of the three operational data sets considered. Because the observed score equating functions are completely determined by the raw score distribution, our simulations are realistic for our purpose. We chose the 2PL model as the data generating model because Haberman (2006) demonstrated that it describes real test data as well as the 3PL model. Although we generate data from an IRT model in order to conveniently manipulate several factors (most importantly, the item difficulties for the anchor tests) in the simulation, we are not fitting an IRT model here. Hence issues of poor IRT model fit are mostly irrelevant to this study.

Simulation Results

The averages of the raw scores on the operational tests are 24.4, 36.7, and 46.6 for the 45-item test, 60-item test, and 78-item test, respectively. The corresponding SD s are 7.1, 10.1, and 12.8, respectively. Because of our simulation design, the means and SD s of the corresponding simulated Y -tests are very close to these values. As in Sinharay and Holland (2006), the average total-test-to-anchor-test correlation is

highest for the miditest followed by the semi-miditest, and then the minitest. For example, for the 78-item test, sample size 5,000, $\Delta_d = .5$, and $\Delta_a = .4$, the averages are .888, .885, and .877, respectively.

Tables 2–7 contain the weighted averages of bias, *SD*, and RMSE, multiplied by 100, for the several simulation conditions.

In these tables, each vertical cell of three values corresponds to a simulation condition. The three numbers in each cell correspond, respectively, to the minitest, the semi-miditest, and the miditest.

Figure 1 shows the equating bias, multiplied by 100, for the CE method for twelve simulation cases with 5,000 examinees and $\Delta_d = .0$.

Each column in the figure corresponds to a value of Δ_a ($-.2, 0, .2, \text{ or } .4$) and each row in the figure corresponds to a value of the number of items (45, 60, or 78). In any plot, the value of $100 \times \text{Bias}(x)$ is shown for all possible values of x (raw score), and two dotted and vertical lines denote the 2.5th percentile and the 97.5th percentile of the true raw score distribution (given by Equation 2) of the test to be equated. The range of the Y axis is the same for all the four plots in any row for convenience of viewing.

The weighted averages of bias, *SD*, and RMSE reported in Tables 2–7 are in units of raw-score points. A difference of .5 or more in the raw score scale is usually a “difference that matters” (DTM), i.e., only a difference more than a DTM leads to different equated raw scores (Dorans & Feigenbaum, 1994). The biases in Tables 2 and 3 never exceed a point, but reach close to a point.

The tables and the figure lead to the following conclusions:

- Effects on bias (Tables 2 and 3; Figure 1):
 - Group difference Δ_a has a substantial effect on bias. Absolute bias is small when $\Delta_a = 0$ and increases as $|\Delta_a|$ increases. This holds for both CE and PSE. This finding agrees with earlier research work such as Hanson and Beguin (2002) and common advice by experts (e.g., Kolen and Brennan, 2004, p. 232) that any group difference leads to equating bias. The sign of Δ_a causes a change in the sign of bias, but does not affect the magnitude of bias.
 - Both CE and PSE are biased, but CE is always less biased than PSE.
 - Anchor test type has a small but nearly consistent effect on bias. Miditests and semi-miditests are usually less biased overall than minitests. This holds for both CE and PSE.
 - “Test length” has a small effect on bias. It is not monotone for PSE and the effect is smaller for CE.
 - Both Δ_d and sample size have almost no effect on bias.
- Effects on *SD* (Tables 4 and 5):
 - Sample size has a large effect on *SD*, which decreases as sample size increases.
 - “Test length” has a modest effect on *SD* for both CE and PSE. *SD* increases as “test length” increases.

TABLE 2
Bias ($\times 100$) for the Different Simulation Conditions: PSE Method

No. of Items	Δ_d	$\Delta_a = -.2$	Number of Examinees												
			100				500				5,000				
			.0	.2	.4	-.2	.0	.2	.4	-.2	.0	.2	.4	-.2	
45	.0	44	00	-.45	-.91	42	-.01	-.45	-.88	43	00	-.43	43	00	-.47
		43	01	-.43	-.87	42	00	-.42	-.84	42	00	-.41	42	00	-.41
		42	01	-.41	-.83	39	00	-.40	-.79	40	00	-.39	40	00	-.39
		42	-.02	-.47	-.91	42	-.01	-.45	-.88	43	00	-.44	43	00	-.44
	.2	41	-.01	-.44	-.88	42	00	-.42	-.84	42	00	-.42	42	00	-.42
		35	-.05	-.46	-.86	39	-.01	-.40	-.79	39	00	-.40	39	00	-.40
		44	-.02	-.47	-.91	42	-.01	-.44	-.88	43	00	-.44	43	00	-.44
		45	-.01	-.45	-.87	42	00	-.42	-.84	42	00	-.42	42	00	-.42
	.5	34	-.04	-.45	-.86	39	-.01	-.40	-.79	39	00	-.40	39	00	-.40
		32	-.04	-.41	-.79	34	-.01	-.36	-.72	35	00	-.35	35	00	-.35
		30	03	-.26	-.56	27	01	-.27	-.56	27	00	-.27	27	00	-.27
		29	01	-.29	-.59	29	02	-.26	-.55	27	00	-.28	27	00	-.28
60	.0	32	-.04	-.41	-.79	34	-.01	-.36	-.72	34	00	-.36	34	00	-.36
		31	03	-.26	-.56	27	00	-.27	-.56	27	00	-.27	27	00	-.27
		29	01	-.28	-.60	29	02	-.26	-.56	27	00	-.28	27	00	-.28
		33	-.04	-.41	-.79	34	-.01	-.36	-.72	34	00	-.36	34	00	-.36
	.2	31	04	-.25	-.56	27	01	-.27	-.56	27	00	-.27	27	00	-.27
		30	01	-.27	-.59	28	02	-.26	-.56	27	00	-.28	27	00	-.28
		35	-.05	-.45	-.86	35	-.03	-.41	-.80	38	00	-.38	38	00	-.38
		33	-.02	-.38	-.74	32	-.01	-.35	-.68	33	00	-.33	33	00	-.33
	.5	30	-.04	-.40	-.75	30	-.03	-.36	-.69	32	00	-.33	32	00	-.33
		36	-.04	-.44	-.86	35	-.03	-.41	-.80	38	00	-.38	38	00	-.38
		34	-.01	-.37	-.75	32	-.01	-.35	-.69	33	00	-.33	33	00	-.33
		31	-.04	-.40	-.76	30	-.03	-.36	-.70	32	00	-.33	32	00	-.33
78	.0	39	-.01	-.40	-.80	38	00	-.38	-.77	38	01	-.38	38	01	-.38
		32	-.03	-.37	-.72	33	00	-.33	-.68	33	00	-.33	33	00	-.33
		35	00	-.33	-.69	33	00	-.32	-.67	33	01	-.32	33	01	-.32
		35	00	-.33	-.69	33	00	-.32	-.67	33	01	-.32	33	01	-.32
	.2	30	-.04	-.40	-.75	30	-.03	-.36	-.69	32	00	-.33	32	00	-.33
		36	-.04	-.44	-.86	35	-.03	-.41	-.80	38	00	-.38	38	00	-.38
		34	-.01	-.37	-.75	32	-.01	-.35	-.69	33	00	-.33	33	00	-.33
		31	-.04	-.40	-.76	30	-.03	-.36	-.70	32	00	-.33	32	00	-.33
	.5	39	-.01	-.40	-.80	38	00	-.38	-.77	38	01	-.38	38	01	-.38
		32	-.03	-.37	-.72	33	00	-.33	-.68	33	00	-.33	33	00	-.33
		35	00	-.33	-.69	33	00	-.32	-.67	33	01	-.32	33	01	-.32
		35	00	-.33	-.69	33	00	-.32	-.67	33	01	-.32	33	01	-.32

Note. The three numbers in each cell correspond, respectively, to the minitest, the semi-midtest, and the midtest.

TABLE 3
Bias ($\times 100$) for the Different Simulation Conditions: CE Method

No. of Items	Δ_d	$\Delta_a = -.2$	Number of Examinees											
			100				500				5,000			
			.0	.2	.4	-.2	.0	.2	.4	-.2	.0	.2	.4	-.2
45	.0	13	-02	-16	-31	12	-01	-15	-29	13	00	-14	-27	
		12	00	-13	-29	12	00	-13	-25	13	00	-12	-24	
		12	00	-13	-26	10	-01	-11	-22	11	00	-11	-21	
		11	-03	-17	-31	12	-02	-15	-29	13	00	-14	-28	
	.2	12	-01	-14	-28	12	00	-13	-25	13	00	-12	-24	
		04	-07	-18	-29	10	-01	-11	-22	11	00	-11	-22	
		14	-03	-17	-30	12	-01	-14	-29	12	00	-14	-28	
		17	-01	-15	-27	13	00	-12	-25	12	00	-12	-25	
	.5	02	-06	-17	-29	10	-01	-10	-22	10	00	-11	-22	
		06	-06	-19	-32	10	-02	-13	-25	11	00	-11	-23	
		09	02	-06	-15	07	00	-07	-15	07	00	-07	-15	
		07	-01	-09	-19	09	02	-06	-15	07	00	-08	-16	
60	.2	06	-06	-18	-32	10	-02	-13	-25	11	00	-12	-24	
		09	03	-06	-15	07	00	-08	-16	07	00	-07	-15	
		08	00	-09	-19	09	02	-06	-15	07	00	-08	-17	
		07	-06	-18	-31	10	-02	-13	-25	11	00	-12	-24	
	.5	10	03	-05	-15	07	00	-08	-16	07	00	-07	-15	
		08	00	-08	-18	09	02	-06	-15	07	00	-08	-17	
		06	-05	-17	-28	07	-03	-14	-24	10	00	-11	-21	
		07	-02	-12	-20	07	-01	-09	-17	08	00	-07	-15	
	78	.2	04	-05	-14	-22	05	-03	-11	-19	07	00	-08	-15
			07	-04	-15	-28	07	-03	-14	-25	11	00	-11	-21
			08	-01	-10	-20	07	-01	-09	-17	08	00	-07	-16
			04	-04	-13	-23	05	-03	-11	-19	07	00	-08	-16
.5		10	-02	-12	-23	11	00	-11	-22	12	01	-11	-22	
		06	-04	-12	-20	08	00	-08	-17	09	01	-08	-17	
		09	00	-08	-17	08	00	-08	-17	09	01	-08	-16	
		09	00	-08	-17	08	00	-08	-17	09	01	-08	-16	

Note. The three numbers in each cell correspond, respectively, to the minitest, the semi-midtest, and the midtest.

TABLE 4
SD ($\times 100$) for the Different Simulation Conditions: PSE Method

No. of Items	Δ_d	$\Delta_a = -.2$	Number of Examinees												
			100				500				5,000				
			.0	.2	.4	.4	-.2	.0	.2	.4	.4	-.2	.0	.2	.4
45	.0	112	112	114	117	51	50	51	52	15	15	15	15	16	
		111	109	111	115	49	48	49	50	15	15	15	15	15	
		109	109	111	114	50	49	50	50	15	15	15	15	15	
		111	110	112	116	51	50	51	52	15	16	15	15	16	
	.2	107	106	107	111	49	48	49	50	15	15	15	15	15	
		108	106	108	111	50	49	50	50	15	15	15	15	15	
		111	111	114	113	51	50	51	51	16	15	15	15	16	
		106	106	107	109	49	49	49	49	15	15	15	15	15	
	.5	106	106	107	109	50	49	50	50	15	15	15	15	15	
		130	129	130	136	56	55	56	57	17	17	17	17	18	
		124	124	125	131	53	53	53	54	17	16	16	17	17	
		123	123	125	131	53	53	53	55	17	16	16	16	17	
60	.2	129	129	130	134	56	55	56	57	17	17	17	17	18	
		123	123	124	129	53	53	53	54	17	16	16	17	17	
		123	122	124	129	53	53	53	55	17	16	16	16	17	
		130	129	129	133	56	55	55	57	17	17	17	17	18	
	.5	124	122	123	128	53	53	53	54	16	16	16	16	17	
		124	122	123	128	53	53	53	54	16	16	16	16	17	
		124	123	124	128	53	53	53	54	16	16	16	16	17	
		157	157	159	169	69	69	69	71	22	21	21	22	22	
	78	.0	158	158	160	168	66	66	67	69	21	21	21	21	21
			158	157	159	167	67	66	67	69	21	21	21	21	22
			147	147	151	159	65	65	65	68	21	20	20	21	21
			149	149	154	161	63	63	64	67	20	20	20	20	21
.5		150	149	154	161	63	63	64	67	20	20	20	21	22	
		149	148	148	166	69	68	68	69	21	21	21	21	21	
		144	143	155	161	68	67	67	68	21	20	20	21	21	
		143	143	154	162	68	67	67	69	21	21	21	21	21	

Note. The three numbers in each cell correspond, respectively, to the minitest, the semi-midtest, and the midtest.

TABLE 5
SD ($\times 100$) for the Different Simulation Conditions: CE Method

No. of Items	Δ_d	Number of Examinees											
		100				500				5,000			
		$\Delta_a = -.2$.0	.2	.4	-2	.0	.2	.4	-2	.0	.2	.4
45	.0	132	131	132	135	59	59	60	60	18	18	18	18
		130	129	129	132	57	56	57	57	18	18	18	18
		128	128	130	132	58	58	58	59	18	18	18	18
		127	127	128	131	59	59	59	60	18	18	18	18
		123	122	123	125	57	56	57	57	18	18	18	18
60	.5	124	122	124	127	58	58	59	59	18	18	18	18
		132	127	128	129	59	58	59	60	18	18	18	18
		120	122	120	124	57	56	57	57	18	18	18	18
		120	122	120	124	58	57	58	58	18	18	18	18
		142	140	141	146	61	60	61	63	19	19	19	19
78	.2	138	137	138	143	59	58	59	60	18	18	18	18
		137	135	138	144	59	59	59	62	18	18	18	18
		142	140	140	144	61	60	61	62	19	19	19	19
		137	137	137	141	59	58	58	60	18	18	18	18
		137	135	137	142	59	59	59	61	18	18	18	18
78	.5	142	140	139	143	61	60	62	62	19	19	19	19
		137	136	136	141	59	58	58	60	18	18	18	18
		138	136	137	141	59	59	59	60	18	18	18	18
		171	172	173	182	77	76	76	78	24	23	23	24
		174	173	174	179	73	73	74	77	23	23	23	23
78	.2	175	173	174	181	75	73	75	77	23	23	23	24
		160	159	162	168	71	71	71	73	22	22	22	23
		161	160	163	167	68	68	69	72	21	21	22	22
		164	162	164	170	69	68	69	72	22	22	22	23
		164	161	175	179	75	74	74	75	23	23	23	24
78	.5	156	155	168	173	73	72	73	75	23	22	23	23
		155	154	167	173	75	73	73	75	23	23	23	24

Note. The three numbers in each cell correspond, respectively, to the minitest, the semi-midtest, and the midtest.

TABLE 6
 RMSE ($\times 100$) for the Different Simulation Conditions: PSE Method

No. of Items	Δ_d	$\Delta_a = -.2$	Number of Examinees											
			100				500				5,000			
			.0	.2	.4	.4	-.2	.0	.2	.4	.4	-.2	.0	.2
45	.0	121	113	123	149	66	50	68	103	46	15	46	88	
		119	111	120	145	64	49	64	98	45	15	44	84	
		117	110	119	142	63	49	64	94	43	15	42	80	
		118	110	121	148	66	50	68	102	45	16	47	88	
	.2	114	106	116	142	64	49	64	98	44	15	45	84	
		113	107	117	141	63	49	64	94	42	15	43	80	
		124	110	123	145	66	50	68	103	46	16	47	89	
		116	106	117	140	64	49	64	98	45	15	45	85	
	.5	112	107	116	139	63	50	64	94	42	15	43	81	
		134	130	137	158	66	55	67	93	39	18	40	74	
		129	124	129	145	61	53	60	80	33	17	33	60	
		128	123	129	147	62	53	60	82	34	17	35	64	
60	.0	134	129	137	157	66	55	67	93	39	18	40	74	
		127	123	128	145	61	53	61	82	33	17	35	64	
		135	129	136	156	65	56	67	93	39	18	41	75	
		128	123	128	142	60	53	61	80	32	17	34	61	
	.2	129	124	129	144	61	54	62	82	33	19	36	65	
		129	124	129	144	61	54	62	82	33	19	36	65	
		161	158	167	191	78	69	81	107	44	22	44	79	
		162	159	166	186	74	67	76	99	40	21	40	71	
	.5	152	147	158	182	74	66	77	100	40	21	40	72	
		154	150	160	181	71	63	74	99	39	20	40	73	
		154	150	161	183	71	63	75	100	39	20	40	73	
		154	148	166	186	79	69	79	105	44	22	44	80	
.5	148	144	161	179	76	67	76	98	39	22	41	73		
	148	144	160	179	76	68	76	98	40	23	41	72		

Note. The three numbers in each cell correspond, respectively, to the minitest, the semi-midtest, and the midtest.

TABLE 7
RMSE ($\times 100$) for the Different Simulation Conditions: CE Method

No. of Items	Δ_d	Number of Examinees											
		100				500				5,000			
		$\Delta_a = -.2$.0	.2	.4	$-.2$.0	.2	.4	$-.2$.0	.2	.4
45	.0	133	131	133	139	60	59	61	67	22	18	23	33
		131	129	130	136	58	56	58	63	22	18	21	30
		129	128	131	136	59	58	59	63	21	18	21	28
		127	127	129	135	60	59	61	67	22	18	23	34
		123	122	124	129	58	56	58	62	22	18	22	30
60	.5	125	123	126	130	59	58	59	63	21	18	21	29
		133	127	129	133	60	59	61	67	23	18	23	34
		121	122	121	128	58	56	57	62	22	18	22	31
		121	123	122	128	59	58	59	63	21	18	21	29
		143	141	143	150	62	61	62	68	23	19	23	31
78	.2	139	138	139	146	60	59	60	63	21	19	21	26
		138	136	139	147	61	59	60	66	22	19	22	30
		143	140	142	148	62	60	62	67	23	19	23	31
		138	138	139	144	60	59	59	63	21	19	21	27
		138	136	138	145	61	59	60	65	22	19	23	31
78	.5	143	141	141	148	62	60	62	68	23	20	23	32
		138	137	138	143	60	59	60	64	22	20	23	29
		139	137	139	144	61	60	61	66	23	21	24	32
		172	173	175	185	77	76	78	82	26	23	26	32
		175	174	176	182	73	73	75	80	25	23	25	30
78	.2	176	174	176	185	75	74	76	81	25	23	25	31
		161	160	163	171	72	71	73	77	25	22	25	31
		162	161	164	170	68	68	70	75	23	22	24	30
		165	162	166	173	70	69	71	76	23	22	24	30
		165	162	176	182	76	74	75	79	27	24	27	33
78	.5	157	156	170	176	75	73	74	78	26	24	26	31
		156	155	168	176	76	74	75	79	27	25	27	32

Note. The three numbers in each cell correspond, respectively, to the minitest, the semi-midtest, and the midtest.

- PSE has slightly less *SD* than CE, especially for small sample size conditions.
 - Δ_a has a small effect on *SD* that is largest for PSE and small sample sizes.
 - Anchor test type has a small effect on *SD* mostly favoring miditest and semi-miditest over minitest, mostly for the small sample size.
 - Δ_d has almost no effect on *SD*.
- Effects on RMSE (Tables 6 and 7; Figure 1):
 - Sample size has a large effect on RMSE, which decreases as sample size increases.
 - Δ_a has a modest effect on RMSE, which increases as $|\Delta_a|$ increases.
 - “Test length” has a modest effect on RMSE. RMSE increases as “test length” increases.
 - CE versus PSE interacts with sample size in its effect on RMSE. PSE is slightly better for the small sample size while CE is much better for the large sample size, and is slightly better for medium sample size.
 - Anchor test type has a small but nearly consistent effect on RMSE favoring miditest and semi-miditest over minitest for both CE and PSE.
 - Δ_d has almost no effect on RMSE.

With respect to the focus of this study, the main conclusion is that the effect of the type of anchor test consistently favors miditests and semi-miditests over minitests,² but is small and not practically significant, and is much smaller than the effects of (a) CE versus PSE for bias, *SD* and RMSE, or (b) sample size on *SD* and RMSE, or (c) Δ_a on bias and RMSE or (d) “test length” on *SD* and RMSE.

The result that CE is better than PSE with respect to equating bias and worse than PSE with respect to *SD* in our simulations augment the recent findings of Wang, Lee, Brennan, and Kolen (2006), who compared the equating performance of PSE and CE in a simulation study. While Wang et al. varied the *SD* of the ability distribution that we did not, we presmoothed the data and varied the sample size and test difficulty difference, something that Wang et al. did not.

Simulations Under NEAT Design from a Multidimensional IRT Model

Simulation Design

We obtained a data set from a licensing test. Of the total 118 multiple choice items in the test, items 1–29 are on language arts, 30–59 are on mathematics, 60–88 are on social studies, and 89–118 are on science. As each of these four content areas can be considered to measure a different dimension, we fitted a four-dimensional IRT model (e.g., Reckase, 1997) with IRF

$$(1 + e^{-(a_1\theta_1 + a_2\theta_2 + a_3\theta_3 + a_4\theta_4 - b_i)})^{-1}, \tag{3}$$

$$\theta = (\theta_1, \theta_2, \theta_3, \theta_4)' \sim \mathcal{N}_4(\mu = (0, 0, 0, 0)', \Sigma),$$

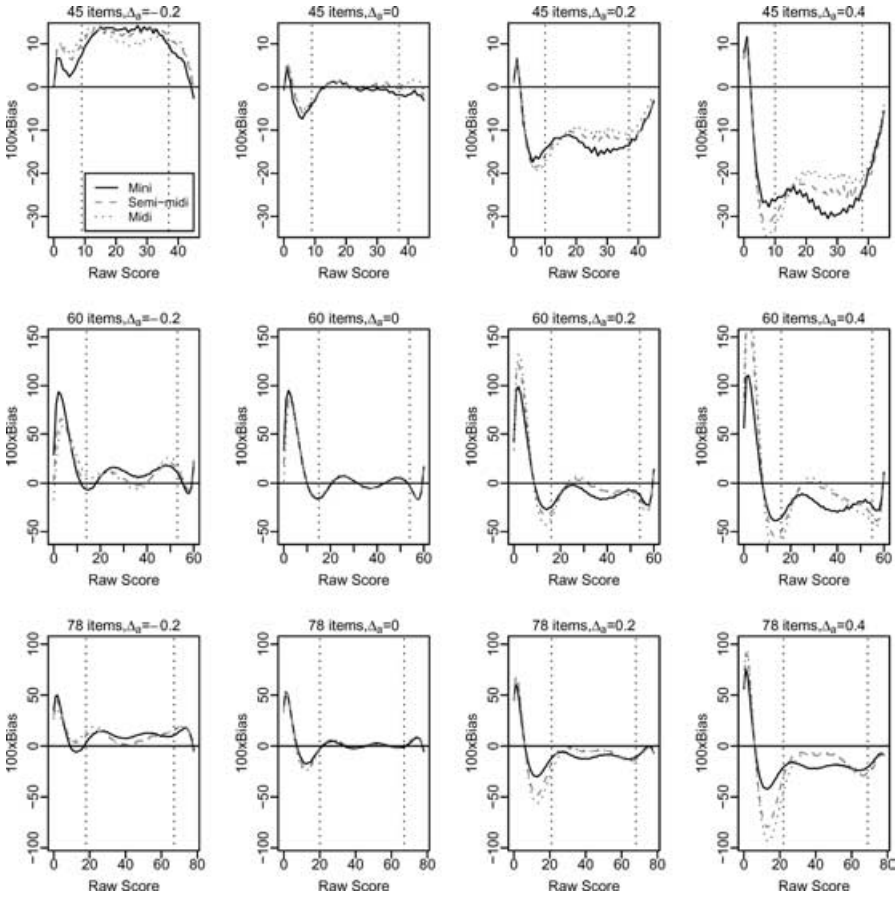


FIGURE 1. Bias (multiplied by 100) in the CE method for tests with 5,000 examinees and $\Delta_a = .0$. The three rows of plots correspond to number of items = 45, 60, and 78. The four columns correspond to population difference $\Delta_a = -.2, 0, .2, \text{ and } .4$.

the symbols having the usual meanings, to the data set. The diagonals of Σ are set to 1 to ensure identifiability of the model parameters. For any item i , only one among a_{1i}, a_{2i}, a_{3i} , and a_{4i} is assumed to be non-zero, depending on the item content (e.g., for an item from the first content area, a_{1i} is nonzero while $a_{2i} = a_{3i} = a_{4i} = 0$), so that we deal with a simple-structure multidimensional IRT (MIRT) model.

The estimated item parameter values were used as generating item parameters of test Y . A bivariate normal distribution \mathcal{D}_k^* was fitted to the log-slope and difficulty parameter estimates corresponding to k th content area, $k = 1, 2, \dots, 4$. The generating item parameters for k th content area for X were randomly drawn from \mathcal{D}_k^* . Because we are generating data from a multidimensional IRT model, X can differ from Y in more complicated ways than for the unidimensional IRT simulation. Hence we manipulated the difficulty parameters of the test X in the following ways to consider several patterns of differences in difficulty between X and Y :

1. No difference (denoted as “N”)—no manipulation.
2. We added Δ_d to the generating difficulty parameters for the first content area for X (denoted as “O” because the difference between X and Y is in *one* dimension).
3. We added Δ_d to the generating difficulty parameters for the first and third content areas for X (denoted as “T” because the difference is in *two* dimensions).
4. We added Δ_d to the generating difficulty parameters of each item in X (denoted as “A” because the difference is in *all* dimensions).
5. We added Δ_d to the generating difficulty parameters for the first and third content areas in X , but subtracted Δ_d from the generating difficulty parameters for the second and fourth content area (denoted as “D” because the difference is *differential* in the dimensions).

We assume that the anchor test has, respectively, 12, 13, 12, and 13 items of the four content areas, leading to an anchor test length of 50. The generating item parameters for the k th content area for the anchor tests were also randomly drawn using the respective distribution \mathcal{D}_k^* . The generating item parameters of the minitest were randomly drawn from the distribution \mathcal{D}_k^* as is. The generating item parameters of the semi-miditest and the miditest were randomly drawn from a distribution that is the same as \mathcal{D}_k^* except for the SD of the difficulties, which was set to one-half and one-tenth, respectively, of the corresponding quantity in \mathcal{D}_k^* . The generating item parameters for the tests X , Y , and the anchor tests were the same for all R replications.

We only used test length of 118 and sample size of 5,000 for the multidimensional simulation. We let Δ_a vary among the three values 0, .2, and .4, and Δ_d among the three values 0, .2, and .5. We used the same three anchor tests (minitest, miditest, and semi-miditest) and the same two equating methods (CE and PSE) as in the unidimensional IRT simulation.

The steps in the simulation are the same as those for the unidimensional IRT simulation except for the following three differences:

- The number of replications is 200 here to reduce computational time.
- The difference between the populations P and Q may be of more complicated nature just like the difference between the tests X and Y . We used $g_Q(\theta) = \mathcal{N}_3(\mathbf{0}, \widehat{\Sigma})$ where $\widehat{\Sigma}$ is the estimate obtained from fitting the model expressed in Equation 3 to the operational test data set. We used $g_P(\theta) = \mathcal{N}_3(\boldsymbol{\mu}_P, \widehat{\Sigma})$, where $\boldsymbol{\mu}_P$, which quantifies the difference between P and Q , was set to be one of the following: 1. $\boldsymbol{\mu}_P = \mathbf{0}$, i.e., no difference (“N”) between P and Q , 2. $\boldsymbol{\mu}_P = (\Delta_a, 0, 0, 0)'$, i.e., is there is difference in one dimension (“O”), 3. $\boldsymbol{\mu}_P = (\Delta_a, 0, \Delta_a, 0)'$, i.e., there is difference in two dimensions (“T”), 4. $\boldsymbol{\mu}_P = (\Delta_a, \Delta_a, \Delta_a, \Delta_a)'$, i.e., the difference is the same in all dimensions (“A”), and 5. $\boldsymbol{\mu}_P = (\Delta_a, -\Delta_a, \Delta_a, -\Delta_a)'$, i.e., differentially different (“D”). The fifth type of difference, “D”, is similar to what was found in, e.g., Klein and Jarjoura (1985; see, e.g., figures 2 and 3 in that paper).
- Application of Equation 2 to compute the true equating function here would have required four-dimensional numerical integration. Hence we take a different approach to compute the true equating function. For each simulation condition, we generate responses to both X and Y of huge examinee samples, of

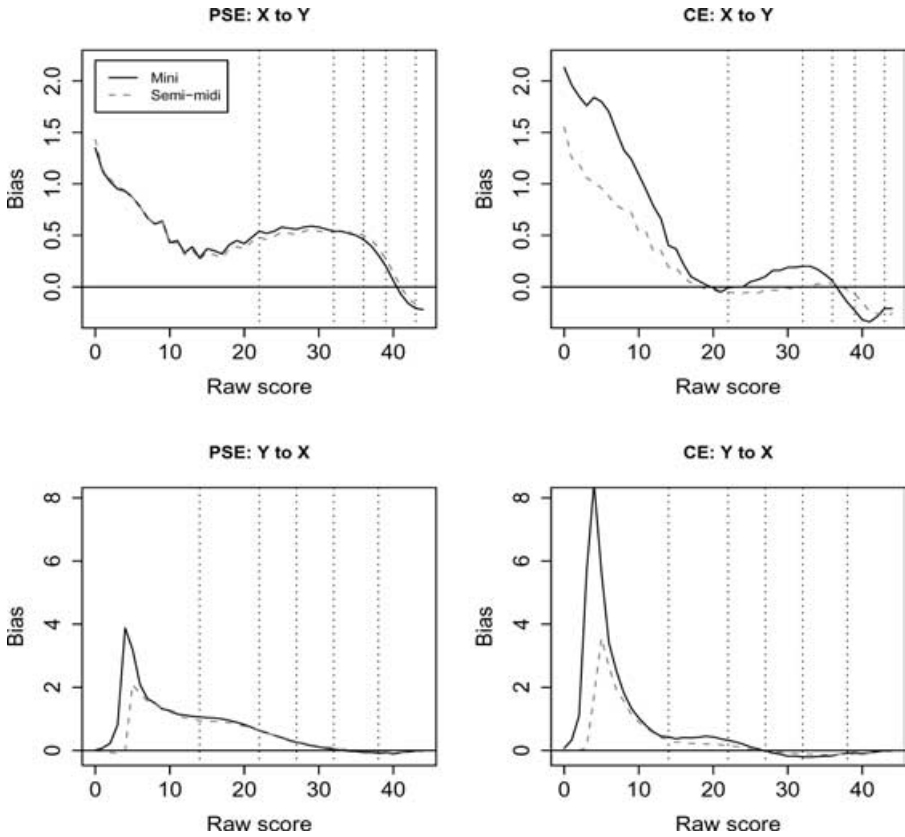


FIGURE 2. Bias of equating for equating X to Y (top row) and Y to X (bottom row) for minitest and semi-midtest for the pseudo-data example.

size 250,000, from P and Q , and perform a single-group equipercentile equating (combining the samples from P and Q) using linear interpolation. We repeated this computation several times with different random seeds—negligible differences between the equating functions obtained from these repetitions ensured that the above method produced the true equating function with sufficient accuracy.

Simulation Results

Tables 8 and 9 show the RMSEs for PSE and CE for the several simulation conditions. Each vertical cell of three values show the RMSEs for a simulation case. The three numbers in each cell correspond, respectively, to the the minitest, the semi-midtest, and the midtest. The SD s (not shown) are very close for all the simulation conditions and the differences in the equating bias (not shown) primarily govern the differences between the RMSEs.

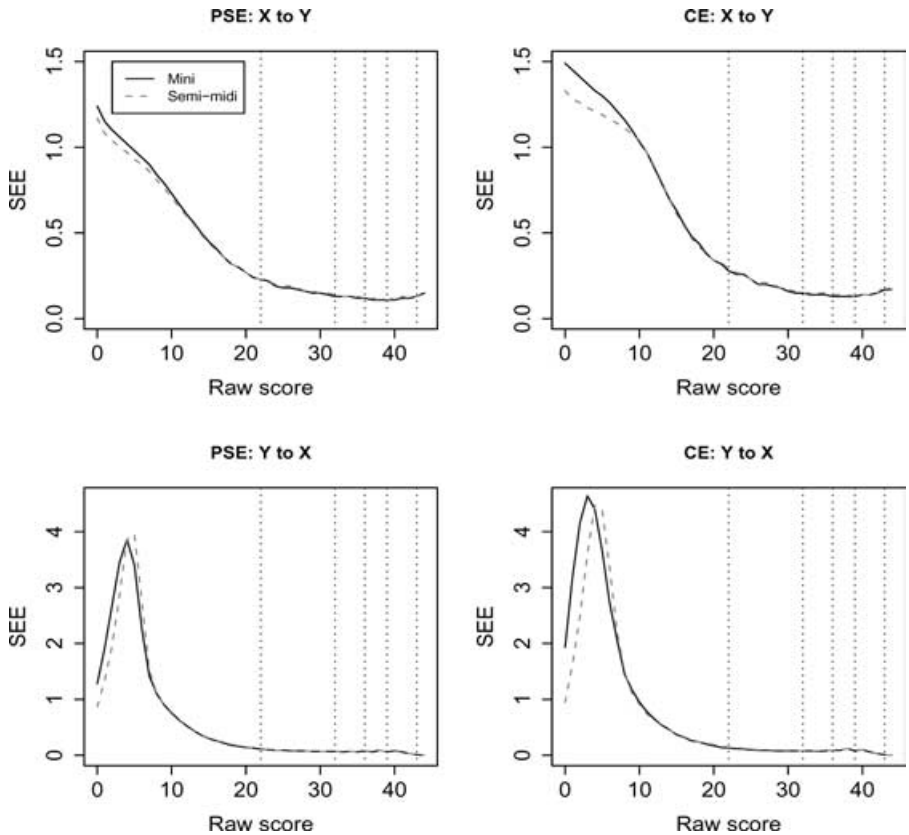


FIGURE 3. *SEE for equating X to Y (top two panels) and Y to X (bottom two panels) for minitest and semi-miditest for the pseudo-data example.*

The factor that has the largest effect on the RMSE is the population difference Δ_a . The RMSE increases as Δ_a increases. The pattern of difference between the two populations also has substantial effect, with pattern “A” associated with the largest values of RMSE. The test differences appear to have no effect on the RMSE.

With respect to the focus of this study, several conditions slightly favor the miditest and semi-miditest while a few others slightly favor the minitest. However, the difference in RMSE between the anchor tests is always small, and far below the DTM, even under conditions (e.g., population difference pattern “D” and $\Delta_a = .4$) that are adverse to equating and worse than what is usually observed operationally. Thus, there seems to be no practically significant difference in equating performance of the three anchor tests.

The results regarding comparison of PSE-vs-CE, which maybe of interest as Wang et al. (2006) did not generate data from a MIRT model, are similar to those from our unidimensional IRT simulation. The RMSE for PSE is mostly larger than CE when the populations are different, the largest differences being observed when population

TABLE 8

RMSE ($\times 100$) for the Different Multidimensional IRT Simulation Conditions for the PSE Method for a 118-Item Test and Sample Size of 5,000 for High-Correlation Case

Population Difference		Test Difference								
		N		O		T		A		D
Pattern	Δ_a	$\Delta_d = 0$.2	.5	.2	.5	.2	.5	.2	.5
N	.0	29	29	30	29	30	30	30	29	30
		28	28	28	28	29	28	30	28	29
		28	28	28	28	29	28	29	28	29
O	.2	31	32	32	32	30	32	31	32	32
		31	31	32	31	30	31	32	31	32
		31	31	34	31	31	31	33	32	35
	.4	37	36	39	34	34	35	36	37	42
		36	37	39	35	35	36	38	37	42
		39	41	44	38	38	39	40	41	48
T	.2	38	39	30	40	32	39	31	42	29
		37	37	31	38	33	37	32	40	29
		37	37	31	38	31	37	32	40	31
	.4	58	35	35	38	41	36	38	33	32
		54	37	36	39	43	37	39	34	32
		56	35	35	36	37	35	37	34	35
A	.2	59	59	56	59	57	59	57	59	56
		58	57	58	58	58	58	59	58	58
		56	56	55	56	56	56	56	56	55
	.4	107	103	103	103	103	103	103	103	103
		105	105	105	106	106	105	106	105	105
		104	098	098	098	098	098	098	098	098
D	.2	29	29	28	29	30	29	30	30	35
		29	29	29	29	31	30	31	29	34
		28	28	30	28	29	28	30	29	40
	.4	30	30	29	34	37	32	32	29	47
		37	33	31	39	42	36	37	30	43
		30	31	32	32	33	30	31	35	56

Note. The three numbers in each cell correspond, respectively, to the minitest, the semi-midtest, and the midtest.

difference is of the type A and $\Delta_a = .4$ (whereas CE leads to RMSEs ranging between .41 and .43, PSE leads to RMSEs ranging between 1.03 and 1.07). Interestingly, the PSE performs slightly better than CE when the population difference is of the type D, even when $\Delta_a = .4$. This finding somewhat contradicts the recommendation of Wang et al. (2006, p. 15) that "... generally speaking, the frequency estimation method does produce more bias than chained equipercntile method and the difference in bias increases as group differences increase" as the "difference in bias" seems to depend in a complicated manner on the type of group difference. This can be a potential future research topic.

TABLE 9
RMSE ($\times 100$) for the Different Multidimensional IRT Simulation Conditions for the CE Method for a 118-Item Test and Sample Size of 5,000 for High-Correlation Case

Population Difference		Test Difference								
		N			O		T		A	
Pattern	Δ_d	$\Delta_d=0$.2	.5	.2	.5	.2	.5	.2	.5
N	.0	32	31	32	31	32	32	32	31	32
		30	30	31	30	31	30	31	30	31
		30	30	30	30	30	30	30	30	30
O	.2	31	32	31	32	31	32	31	32	32
		30	31	31	31	31	30	31	30	31
		30	31	32	31	31	30	31	31	33
	.4	32	31	33	31	31	31	32	32	35
		31	32	33	32	32	32	32	32	34
		33	34	36	32	32	33	33	34	39
T	.2	32	32	31	33	31	33	31	34	31
		31	31	31	32	31	31	31	32	30
		31	31	32	31	31	31	32	32	33
	.4	36	31	31	32	34	32	32	31	32
		34	31	31	32	34	32	32	31	31
		35	35	36	34	33	35	35	36	40
A	.2	34	34	33	34	34	34	34	34	33
		33	33	34	34	34	34	34	34	34
		33	32	33	33	33	33	33	32	33
	.4	43	42	42	42	42	42	43	41	41
		42	43	43	43	43	43	44	43	42
		41	40	39	40	40	39	40	39	39
D	.2	31	31	31	32	31	31	31	34	40
		31	31	31	31	32	32	32	31	37
		30	30	34	31	31	30	32	32	45
	.4	32	32	34	32	34	32	34	37	60
		34	32	32	34	37	34	35	33	53
		32	37	41	34	33	35	36	46	71

Note. The three numbers in each cell correspond, respectively, to the minitest, the semi-miditest, and the miditest.

For the data set from the 118-item test, the estimated correlations between the components of θ range between .73 and .89, which can be considered too high for the test to be truly multidimensional. Hence, we repeated the simulations by considering a variance matrix (between the components of θ) Σ^* whose diagonals are the same as those of Σ , but whose off-diagonals are re-adjusted to make each correlation implied by Σ^* .15 less than that implied by Σ . This brings down the correlations to values (between .58 and .74) that are high enough to be practical, but also low enough for the test to be truly multidimensional. The results for these simulations are similar to those in Tables 8 and 9 and are not reported. We also considered several content

nonrepresentative anchor tests, but they had mostly large RMSEs (results not shown), demonstrating the importance of content representativeness of the anchor tests. We also repeated the above multidimensional IRT simulation procedure using an admissions test data set; we fitted a 3-dimensional MIRT model as the test has three distinct item types; the results (not shown) were similar as above, i.e., there was hardly any difference in equating performance of the three types of anchor tests.

Pseudo-Data Example

It is not easy to compare a minitest versus a miditest or semi-miditest in operational setting, as almost all operational anchor tests are constructed to be minitests. However, a study by von Davier, Holland, and Livingston (2005) allowed us perform the comparison, even though it is rather limited because of short test lengths and short anchor lengths. The study considered a 120-item test given to two different examinee samples P and Q of sizes 6,168 and 4,237, respectively. The sample Q has a higher average score, by about a quarter in SD -of-raw-score unit. Two 44-item tests X and Y , as well as anchor tests (that were constructed to be minitests) of lengths 16, 20, and 24 were constructed by partitioning the 120-item test. The 20-item anchor was a subset of the 24-item anchor and the 16-item anchor was a subset of the 20-item anchor. The test X was designed to be much easier (the difference being about 128% in SD -of-raw-score unit) than the test Y .

Of the total 120 items in the test, items 1–30 are on language arts, 31–60 are on mathematics, 61–90 are on social studies, and 91–120 are on science. As the “minitest,” we take the 16-item anchor test of von Davier et al. (2005). There were not enough middle-difficulty items to choose a miditest. The semi-miditest we chose was a subset of the 24-item anchor test of von Davier et al. We ranked the six items within each of the four content areas in the 24-item anchor test according to their difficulty (proportion correct); the four items ranked 2nd to 5th within each content area were included in the 16-item semi-miditest. Nine items belonged to both the minitest and semi-miditest. We refer to this example as a “pseudo-data” example rather than a “real data” example because the total tests and the anchor tests we consider were not operational, but artificially constructed from real data.

Note that by construction, the semi-miditest, like the minitest, is content representative. Also, the semi-miditest has roughly the same average difficulty as the minitest; the average difficulties of the minitest and the semi-miditest are .68 and .69, respectively, in P , and .72 and .73, respectively, in Q . However, the spread of the item difficulties of the semi-miditest is less than that of the minitest. For example, the SD of the item difficulties of the minitest and the semi-miditest are .13 and .09, respectively, in P , and .12 and .08 in Q (the SD of the item difficulties for X in P is .12 while that for Y in Q is .17).

The first four rows of Table 10 show the relevant anchor-test-to-total-test correlation coefficients.

We computed the equating functions for PSE and CE equipercentile methods, using presmoothing and linear interpolation, for the minitest and the semi-miditest for equating X to Y by pretending that scores on X were not observed in Q and scores on Y were not observed in P (i.e., treating the scores on X in Q and on Y in P as

TABLE 10
Findings from the Long Basic Skills Test

	Minitest	Semi-Miditest
Correlation for X and A in P	.75	.73
Correlation for Y and A in Q	.73	.68
Correlation for X and A in Q	.76	.73
Correlation for Y and A in P	.71	.68
Weighted average of bias: Equating X to Y , PSE	.31	.34
Weighted average of absolute bias: Equating X to Y , PSE	.36	.37
Weighted average of bias: Equating X to Y , CE	-.05	-.06
Weighted average of absolute bias: Equating X to Y , CE	.18	.08
Weighted average of bias: Equating Y to X , PSE	.34	.35
Weighted average of absolute bias: Equating Y to X , PSE	.36	.36
Weighted average of bias: Equating Y to X , CE	.06	.03
Weighted average of absolute bias: Equating Y to X , CE	.21	.12

missing), and then for equating Y to X by pretending that scores on Y were not observed in Q and scores on X were not observed in P . We also computed the criterion (“true”) equating function by employing a single-group equipercentile equating with linear interpolation on all the data from the combined sample of P and Q .

Figure 2 shows a plot of the bias in equating X to Y and Y to X for the semi-miditest and minitest. The bias here is defined as the difference between an equating function and the above-mentioned criterion equating function. Each panel of the figure also shows using vertical lines the five quantiles, for $p = .025, .25, .50, .75, .975$, of the scores on the test to be equated in the combined sample including P and Q .

Figure 3 shows a plot of the SEE for equating X to Y and for equating Y to X for the semi-miditest and minitest. Contrary to the intuitive expectations of some of our colleagues, Figure 3 shows that at the extreme scores the SEE obtained by using a mini anchor test is not smaller than the SEE obtained by using a semi-miditest. If anything, the opposite phenomenon holds. Also, Figures 1 and 3 show that at the extreme scores the bias for a semi-miditest is not systematically worse than that for a minitest either.

Table 10 also shows weighted averages of equating bias, the weight at any score point being proportional to the corresponding frequency in the combined sample.

There is hardly any difference between the minitest and the semi-miditest with respect to equating bias and SEE, especially in the region where most of the observations lie. The PSE method slightly favors the minitest while the CE method slightly favors the semi-miditest. Compared to the PSE method, the CE method has substantially lower equating bias and marginally higher SEE.

Thus, the pseudo-data example, even with its limitations, such as short test and anchor test lengths and large difference between the total tests, provides us with some evidence that a semi-miditest does not perform any worse than a minitest in operational equating.

Discussion and Conclusions

This article examines the choice of anchor tests for observed score equating, and challenges the traditional view that a “minitest” is the best choice for an anchor test. Several simulation studies and a pseudo-data example are used to study the equating performance, especially equating bias and the SEE, of several anchor tests, including those having statistical characteristics that differ from those of a minitest. We show 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 minitests in equating. Note that our suggested anchor tests will often be easier to construct operationally than minitests.

Thus, our results suggest that the requirement of an anchor test to have the same spread of item difficulty as the total tests may be too restrictive and need not be optimal. The design of anchor tests can be more flexible than the use of minitests without losing any important statistical features in the equating process. Our recommendation then is to enforce a restriction on the spread of item difficulties of an anchor test only when it leads to operational convenience. For example, for tests using internal anchors, using a minitest (i.e., restricting the spread to be the same as that of the total tests) may be more convenient because the scarce extreme difficulty items can be used in the anchor test and hence in both of the tests to be equated. For external anchors, our recommendation is to worry about content, average difficulty, and any other requirement, but not about spread of difficulty.

Our findings will be most applicable to testing programs using external anchors. All the results reported in this article were obtained using external anchors. Though some limited simulations (results not reported) showed that the miditest and semi-miditest perform as well as the minitest even for internal anchors, we do not recommend the use of the formers to the internal anchor case because of the above-mentioned reason (the scarce items being used in both of the tests to be equated) and also because including middle difficulty items in the anchor test might create difficulties for the test developers to meet the test specifications when they choose the remaining items in the total test.

Though not the focus of the article, we also find interesting results regarding the comparison of PSE and CE that augment the recent findings of Wang et al. (2006). Both of these studies find that CE has less equating bias and more SEE than PSE in general. However, our work is more extensive than Wang et al. regarding some aspects, e.g., we simulate data under a MIRT model (that can be argued to reflect reality better than a unidimensional IRT model) and perform presmoothing of the data using loglinear models.

Before full-scale operational use of miditests and semi-miditests, the following issues may need attention:

1. A comparison of the performance of miditests and semi-miditests with minitests under more conditions is needed. For example, this study did not vary factors such as
 - (a) ratio of the length of the anchor and the total test given a total test length,
 - (b) mean difficulty of the anchor test in comparison to that of the total tests,

- (c) distribution of the item difficulties for the total test,
 - (d) ratio of the sizes of the samples from P and Q ,
 - (e) the SD of the generating ability distributions, and
 - (f) the difference in mean ability of Q and mean difficulty of Y . The difference was set equal in our simulations; however, the semi-midtest and the midtest performed as well as the minitest in limited simulations performed by setting these two quantities unequal (results not reported here).
2. A comparison of midtests and semi-midtests with minitests for several operational data sets should be performed.
 3. It will be useful to examine the equating performance of midtests and semi-midtests for other types of equating methods such as IRT true score equating. It may be useful to consider other equating criteria like the same distributions property (Kolen & Brennan, 2004) and the first- and second-order equity property (e.g., Tong & Kolen, 2005).
 4. The effect of midtests and semi-midtests on the robustness of anchor tests to varying degrees of differential item functioning (DIF) should be examined. It may happen that the difficulty of some anchor test items changes between the two test administrations because of context effects, test security issues, etc.—and it will be useful to examine whether minitests or midtests are more robust to such problems.
 5. The following practical issues should be considered:
 - (a) When the anchor test is external, can the examinees easily find it if it is a midtest (and be less motivated to answer that)?
 - (b) How to choose a midtest or semi-midtest when the anchor items are mostly based on a shared stimulus like a reading passage.

Notes

¹A content-representative anchor is one in which the proportion of items in each content area is the same as, or very similar to, that of the total tests.

²The only simulation cases when the minitest performs better compared to midtest and semi-midtest correspond to 78 items, 100 examinees, and $\Delta_d = 0$ or $.2$ (e.g., see Tables 6 and 7); even for these cases, it was found that the advantage of the minitest is not statistically significant.

Acknowledgments

The authors thank Michael Kolen, Alina von Davier, Shelby Haberman, Neil Dorans, Samuel Livingston, Tim Moses, Dan Eignor, Gautam Puhan, Jinghua Liu, and the reviewers for their invaluable advice. The authors gratefully acknowledge the help of Kim Fryer with proofreading. Any opinions expressed in this publication are those of the authors and not necessarily of Educational Testing Service.

References

- Angoff, W. H. (1968). How we calibrate College Board scores. *College Board Review*, 68, 11–14.
- Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

- Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. *Journal of Educational Measurement*, 22(1), 13–20.
- 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, 225–244.
- Davey, T., Nering, M. L., & Thompson, T. (1997). *Realistic simulation of item response data (Research Report 97-4)*. Iowa City, IA: ACT, Inc.
- Dorans, N. J., & Feigenbaum, M. D. (1994). *Equating issues engendered by changes to the SAT and PSAT/NMSQT (ETS Research Memorandum 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 SR-98-02)*. Princeton, NJ: ETS.
- Haberman, S. J. (2006). *An elementary test of the normal 2PL model against the normal 3PL model (ETS RR-06-10)*. Princeton, NJ: ETS.
- Hanson, B. A., & Beguin, A. A. (2002). Obtaining a common scale for the item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, 26, 3–24.
- Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, 25(2), 133–183.
- Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with non-random groups. *Journal of Educational Measurement*, 22, 197–206.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.
- Lord, F. M. (1950). *Notes on comparable scales for test scores (RB-50-48)*. Princeton, NJ: Educational Testing Service.
- Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score “equatings.” *Applied Psychological Measurement*, 8, 453–461.
- Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating method. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 71–135). New York: Academic Press.
- Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). Washington, DC: American Council on Education.
- Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). Hillsdale, NJ: Erlbaum.
- Sinharay, S., & Holland, P. W. (2006). *The correlation between the scores of a test and an anchor test (ETS RR-06-04)*. Princeton, NJ: Educational Testing Service.
- Tong, Y., & Kolen, M. J. (2005). Assessing equating results on different equating criteria. *Applied Psychological Measurement*, 29(6), 418–432.
- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of equating*. New York: Springer.
- von Davier, A. A., Holland, P. W., & Livingston, S. A. (2005). An evaluation of the kernel equating method: A special study with pseudo-tests from real test data. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.
- Wang, T., Lee, W., Brennan, R. L., & Kolen, M. J. (2006). A comparison of the frequency estimation and chained equipercentile methods under the common-item non-equivalent groups design. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Authors

SANDIP SINHARAY is a Senior Research Scientist, Educational Testing Service, MS 12T, Rosedale Road, Princeton NJ 08541; ssinharay@ets.org. His primary research interests include equating, differential item functioning, item response theory, Bayesian methods, and statistical computing.

PAUL W. HOLLAND is a Consultant, Educational Testing Service, MS 12T, Rosedale Road, Princeton NJ 08541; pholland@ets.org. His primary research interests include application of statistics to social science research, discrete data analysis, test linking, differential item functioning, item response theory, and causal inference.