



*Research
Report*

The Correlation Between the Scores of a Test and an Anchor Test

Sandip Sinharay

Paul Holland

The Correlation Between the Scores of a Test and an Anchor Test

Sandip Sinharay and Paul Holland
ETS, Princeton, NJ

March 2006

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

<http://www.ets.org/research/contact.html>

Copyright © 2006 by Educational Testing Service. All rights reserved.

ETS and the ETS logo are registered trademarks of
Educational Testing Service.



Abstract

It is a widely held belief that an anchor test used in equating should be a miniature version (or *minitest*) of the tests to be equated; that is, the anchor test should be proportionally representative of the two tests in content and statistical characteristics. This paper examines the scientific foundation of this belief, especially regarding statistical characteristics. In particular, we study, through several simulation studies and a real data example, the correlation of the raw score of the total test with several anchor tests, including those having statistical characteristics that differ from those of the total test. Anchor tests with spread of item difficulties less than that of a total test are shown to consistently have higher correlations than a minitest. Because of the general view that higher correlations of the scores of an anchor test and the total test lead to better equating, our results suggest that requiring an anchor test to mimic the statistical characteristics of the total test may be too restrictive and may not be optimal.

Key words: Equating, minitest, NEAT design

1. Introduction

The use of an *anchor test* is very common in test equating (see, e.g., Angoff, 1971; Kolen & Brennan, 2004; Livingston, 2004; Michaelides & Haertel, 2004; Petersen, Kolen, & Hoover, 1989; Petersen, Marco, & Stewart, 1982). The anchor test is crucial to the quality of equating in a nonequivalent group anchor test (NEAT) design, which is the most popular equating design used operationally.

It is a widely held belief that the anchor test should be a *mini version* of the total test. Angoff (1968, p. 12) and Budescu (1985, p. 15) recommended an anchor test that is a *parallel miniature of the operational forms*. More specifically, several experts recommended that an anchor test should be proportionally representative or a mirror of the total test in content and statistical characteristics (von Davier, Holland, & Thayer, 2004, p. 33; Dorans, Kubiak, & Melican, 1998, p. 3; Kolen & Brennan, 2004, p. 19; Petersen, Kolen, & Hoover, 1989, p. 246). Currently, most testing programs employ this type of anchor test (referred to as a *minitest*).

The requirement that the anchor test should be representative of the total test with respect to content is justified from the perspective of content validity and has been shown to be important by Klein and Jarjoura (1985) and Cook and Petersen (1987). Peterson et al. (1982) demonstrated the poor performance in equating of anchor tests whose mean difficulty are different from that of the total tests. However, there is no rigorous argument or proof (to our knowledge) as to why the spread of the difficulty of the anchor test should be representative of the total test. Furthermore, this requirement often makes the creation of a minitest more time-consuming. That is because there is a need to include very difficult or very easy items in the minitest to make its spread of difficulty the same as that of the total test and that such items are scarce. An anchor test that relaxes the requirement on the spread of the item difficulties will be more convenient operationally.

Interestingly, a study of the psychometrics literature on the relationship of item difficulty and characteristics of a test does not make it obvious that a minitest is the ideal anchor test. For example, Gulliksen (1945, p. 79) commented that in order to maximize the reliability and variance of a test, the items should have high intercorrelations, all items should be of the *same difficulty level*, and the level should be as near to 50% as possible.

Richardson (1936) commented that “the present investigation confirms the conclusion of Thelma Gwinn Thurstone that a test composed of items of 50 per cent difficulty has a general validity which is higher than tests composed of items of any other degree of difficulty,” where the term *validity* refers to the correlation of the scores on a test with those on a criterion test. Cronbach and Warrington (1952, p. 127) found that for items of the type ordinarily used in psychological tests, a test with *constant item difficulty* gives greater overall validity and superior validity of most cutting scores, compared to a test with a range of item difficulties. Dealing with tests where guessing occurs, Lord (1952) found that reliability and the curvilinear correlation of the test score on criterion (that is shown to be equal to the square root of the test reliability), which he thought was a more appropriate measure to examine than the simple correlation, will be maximized by (a) *minimizing the variability of item difficulty*, and (b) making the level of item difficulty somewhat easier than the halfway point between a chance percentage of correct answers and 100

The above mentioned comments were made in the context of total tests, but they hold equally well for anchor tests. Thus, in the light of the literature, it is apparent that requiring an anchor test to have statistical characteristics similar to that of the total test may be too restrictive, and that anchor tests with spread of item difficulties less than that of the total test might perform equally well. For example, consider an anchor test that is representative of the total test with respect to content, but is on the other extreme from the minitest with regard to item difficulty; that is, it consists only of items of medium difficulty. This anchor test will be referred to as a *miditest* henceforth. According to the above mentioned references, the miditest will have higher reliability and, if the total test is assumed to be the criterion, a higher correlation with the total test (which is one criterion of a good anchor test, as described below) than will a minitest. An anchor test with spread of item difficulties less than that of a minitest (or total test) but more than that of a miditest (called a *semi-miditest* henceforth) might be of interest as well; while a perfect miditest will be difficult to obtain operationally, a semi-miditest will not (conceptually, a semi-miditest can be obtained by replacing a few extreme-difficulty items from an operational minitest by medium-difficulty items).

The natural next step then is to study, in actual equating conditions, the performance

of anchor tests with statistical characteristics that differ from those of the total test compared to that of the minitest, which, unfortunately, is not straightforward. Almost all the anchor tests used operationally are minitests, so that to perform such a study, one has to collect new data on operational tests. Convincing test administrators to operationally employ an anchor test with statistical characteristics that differ from those of the total test may not be straightforward, as such anchor tests are contrary to the common belief; there needs to be some preliminary results in support of such anchors that will persuade test administrators to consider employing it. Even if one performs equating by operationally implementing a minitest and one of the suggested anchor tests, there is not a clear criterion to compare their performances against. For example, computation of bias in equating, as it involves knowing the true results, is a difficult proposition—one needs a very specially designed study (to be described in detail in the last section of this report) to compute bias. We overcome the above mentioned difficulties by choosing a criterion, the anchor-test to total-test correlation, that is very closely related to equating, yet will allow us to systematically compare the minitest and the anchor tests with statistical characteristics that differ from those of the total test.

The role of correlation in equating is important. The correlation coefficient is expected to be inversely related to the standard error of equating (SEE). Let us consider the simple case of the random groups with anchor test design (Angoff, 1971; Lord, 1950) in which randomly equivalent groups of examinees are administered two forms that contain common items (i.e., an anchor test). Suppose X and Y are the tests to be equated, and suppose A denotes the anchor. Under assumptions that (a) that the populations taking tests X and Y are randomly equivalent, (b) scores on X and A are bivariate normally distributed, (c) scores on Y and A are bivariate normally distributed, (d) the correlation between scores on X and A is equal to that between scores on Y and A , and (e) the sample sizes for examinees taking the old and new forms are equal, the square of the SEE at any value x_i of score in test X is approximated by

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

where symbols have usual meanings. Equation 1 shows that as the anchor-test to total-test

correlation ρ_{XA} increases, the standard error of equating decreases (the same phenomenon mentioned by Budescu, 1985, p. 15). As a result, in this case, the higher correlation for the miditest will result in a lower SEE.

Further, the anchor test score plays the role of a covariate in Tucker equating and frequency estimation equipercentile equating (e.g., Kolen & Brennan, 2004) and it is well-known that the performance of a covariate regarding bias removal and variance reduction improves as its correlation with the response variable (the total test score here) increases.

Furthermore, several experts have suggested that a higher correlation of an anchor test with the total tests leads to better equating. Angoff (1971, p. 577) commented that the usefulness of an anchor test depends on the extent to which it is correlated with the tests being equated. Dorans et al. (1998, pp. 2-3) recommended that

prescriptions about number of items, content, and statistical representation are means of achieving an end: namely, scores on an equating test that have a very high statistical relationship (often measured by a correlation) with scores on each of the tests to be equated.

Budescu (1985, p. 15) commented that

a high correlation between U (anchor test) and X (total test) is a necessary condition for an efficient linear equating of the two forms. Therefore, this correlation, or some monotonic transformation of it, could be used as a global measure of the efficiency of the equating. . . . To achieve a high correlation, special effort and care should be put into selecting the items of the anchor test and incorporating them into the operational forms (e.g., Angoff, 1968).

Petersen et al. (1989, p. 246) and von Davier et al. (2004, p. 33) suggested that the higher the correlation between scores of an anchor test and scores on the tests to be equated, the better the anchor test is for equating.

Therefore, as a preliminary step in comparing a minitest and anchor tests with statistical characteristics that differ from those of the total test, this paper examines the

correlation coefficients between the scores of a total test and several anchor tests through a series of simulations and real data examples. There are several other requirements of a good anchor test that are not our focus here (see, for example, Angoff, 1971; Kolen & Brennan, 2004; Petersen et al., 1989).

Section 2 performs detailed simulation studies, generating data from both univariate and multivariate item response theory (IRT) models, to compare, with respect to anchor-test to total-test correlation, the minitest and anchor tests with statistical characteristics that differ from those of the total test. Section 3 describes a real data example. Section 4 provides discussion and future work.

2. Simulation Studies

We performed a series of simulations to study the anchor-test to total-test correlation for different types of anchor tests, including those with statistical characteristics that differ from those of the total test.

2.1 Simulations for Univariate Tests

The simulations described in this section explore the correlation between scores on a total test and an external anchor test, generating data from various univariate IRT models. Note that even though we are generating data from the IRT models, there is no IRT equating involved in this paper.

2.11 Study Design

Four factors were involved in these simulations. First, data were generated from the Rasch model or the 2PL model. For the 2PL model, the logarithm of the generating slope parameters were randomly drawn from a normal distribution with mean 0 and standard deviation (SD) 0.5.

Second, the difficulty parameters of the total test were one of the two following types:

- equispaced between -2 and 2; this is called *equispaced total test*. Most operational tests are of this type.

- all fixed at 0; this is called *constant difficulty total test*. Mastery tests usually are of this type.

Third, the difficulty parameters of the anchor test were set according to one of the following schemes:

- equispaced between -2 and 2. This corresponds to the minitest.
- all fixed at 0. This corresponds to the miditest.
- equispaced between -1 and 1. This corresponds to the semi-miditest. Note that the spread/SD of the difficulty parameters for this test lies between that of the miditest and the minitest.

Note that for the equispaced total test, the last two of the above three anchor tests have statistical characteristics that differ from those of the total test.

The fourth factor studied is $p(\theta)$, the generating distribution of the proficiency θ , which was assumed to be either $\mathcal{N}(0, 1)$ (average) or $\mathcal{N}(1, 1)$ (high). The correlations are expected to be higher when the distribution of the item difficulty parameters matches $p(\theta)$.

The data generator generated 100 data sets with 1,000 examinees for each combination of IRT model, total test type, anchor test type, and ability distribution type. The total test had 40 items, while each anchor test had 20 items.

2.12 Results

Table 1 reports the average and SD of the 100 correlation coefficients for the $2 \times 2 \times 3 \times 2 \equiv 24$ cases.

The standard error for each average correlation can be obtained by dividing the corresponding SD by 10.

Figure 1, showing a graphical plot of the values of the correlation, makes the comparison between the different cases convenient.

Different plotting symbols (as shown in the legend) as well as different connecting lines are used for the different types of anchor tests.

Figure 2 shows the main effects for the four different factors, plotting the average correlations for each factor after averaging over the other factors.

Table 1.
The Average (and SD, Within Parentheses) of the Correlation Coefficients for Data Generated From the Unidimensional IRT Models

	Data generating model							
	Rasch model				2PL model			
	Total–equispaced		Total–constant		Total–equispaced		Total–constant	
Anchor	$\mathcal{N}(0, 1)$	$\mathcal{N}(1, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(1, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(1, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(1, 1)$
Mini-test	0.798 (0.010)	0.783 (.013)	0.807 (0.010)	0.790 (0.011)	0.792 (0.020)	0.779 (0.020)	0.803 (0.020)	0.782 (0.020)
Semi-miditest	0.812 (0.009)	0.797 (.011)	0.822 (0.009)	0.806 (0.012)	0.811 (0.017)	0.791 (0.019)	0.828 (0.018)	0.804 (0.020)
Midi-test	0.815 (0.010)	0.799 (.012)	0.826 (0.009)	0.812 (0.012)	0.815 (0.017)	0.793 (0.019)	0.834 (0.017)	0.811 (0.020)

Figure 3 shows the first-order interaction effects between anchor test type and each of the other three factors, plotting the average correlations (averaging over the other two factors).

The captions of these figures provide further descriptions of them. All the effects shown in these two figures were found significant by an analysis of variance (ANOVA) of the correlations shown in Table 1. However, the nearly parallel lines in each graph indicate that the effect of anchor test type is nearly the same for the levels of the other factors.

The main findings from Table 1 and Figures 1 to 3 are:

1. The miditest has the highest average correlation for all the situations considered.
2. The semi-miditest performs almost as well as the miditest with respect to correlation. This is an important finding because operationally, it may be impossible to obtain a miditest, but it should not be too difficult to produce a semi-miditest.
3. Both the miditest and the semi-miditest always have higher average correlations than the minitest does with the total test. The difference is statistically significant in all cases.

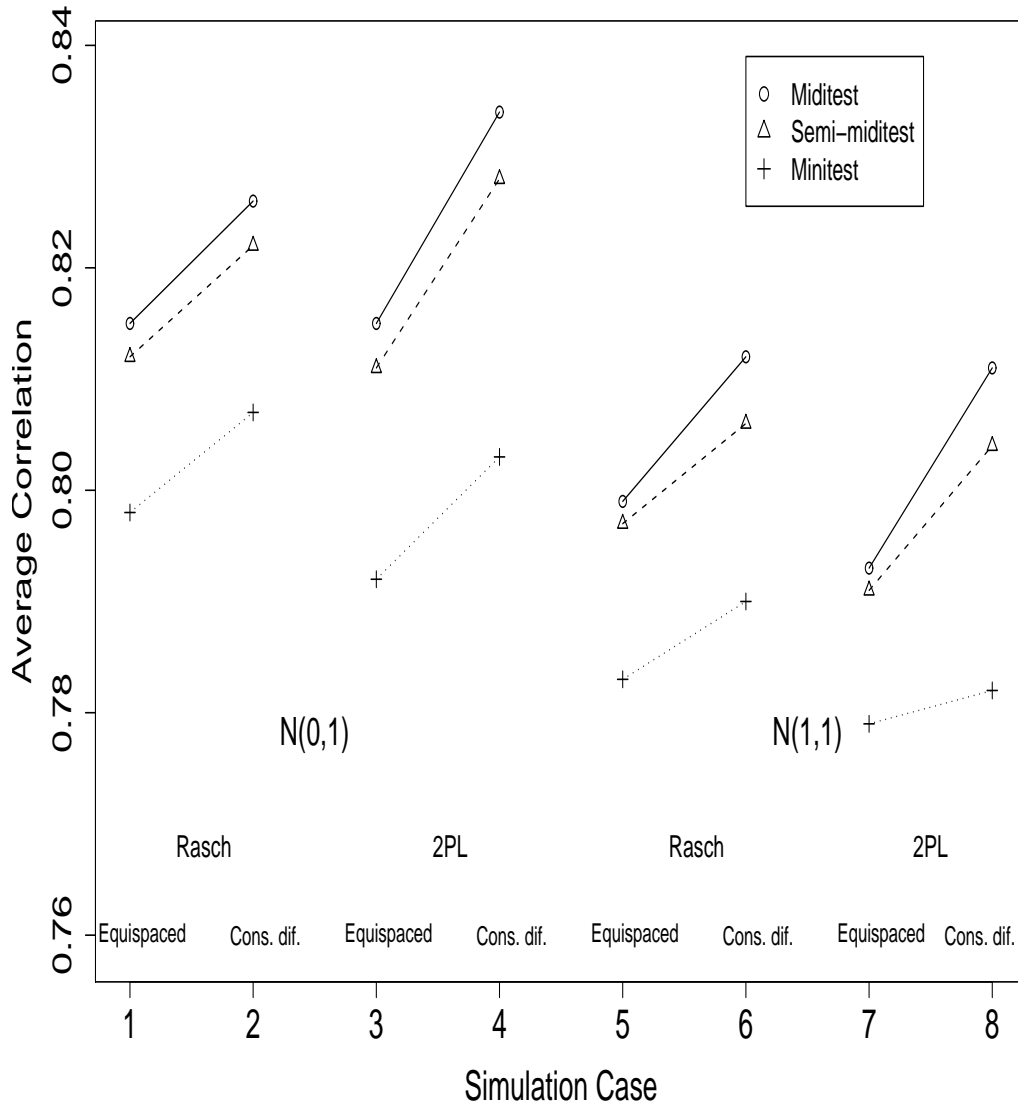


Figure 1. Plot of the average correlation coefficients for the different cases in the univariate IRT simulation study.

Note. The left-most four cases correspond to $p(\theta)$ of $N(0,1)$, while the right-most four cases correspond to $p(\theta)$ of $N(1,1)$. Out of the four cases for any $p(\theta)$, the left-most two correspond to the Rasch model while the right-most two correspond to the 2PL model. Out of the pair of cases for any combination of $p(\theta)$ and IRT model, the equispaced total test case (denoted *Equispaced*) is plotted to the left, while the constant difficulty total test case (denoted *Cons. dif.*) is plotted to the right.

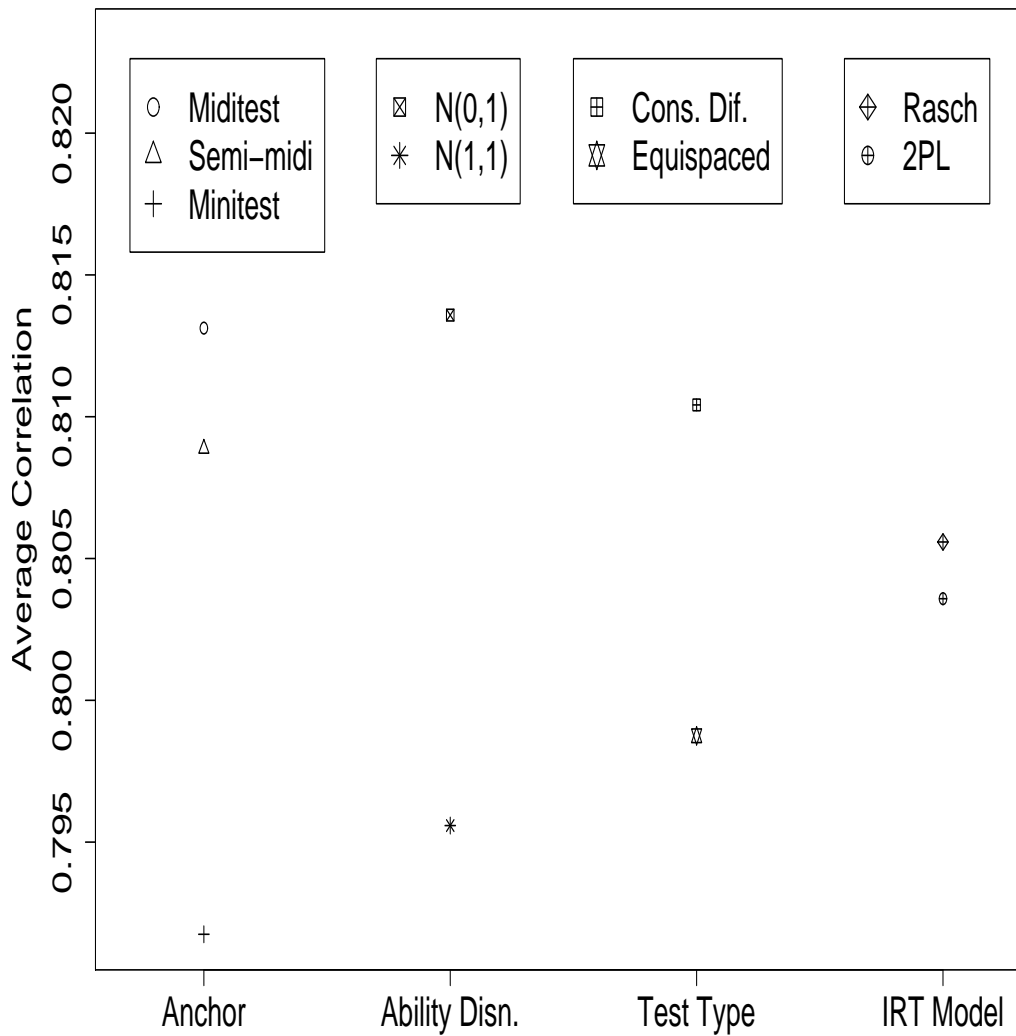


Figure 2. Plot showing the main effects for the different factors in the simulation studies for univariate IRT models.

Note. The labels for the horizontal axis denote the main effects plotted. At the extreme left, average correlations for the three anchor tests (averaging over the other three factors, ability distribution, IRT model, and total test type) are shown; then, the average correlations for the two ability distributions are shown, followed by those for the two test types and the two IRT models. The plotting symbols are shown for each factor, above the plotted points.

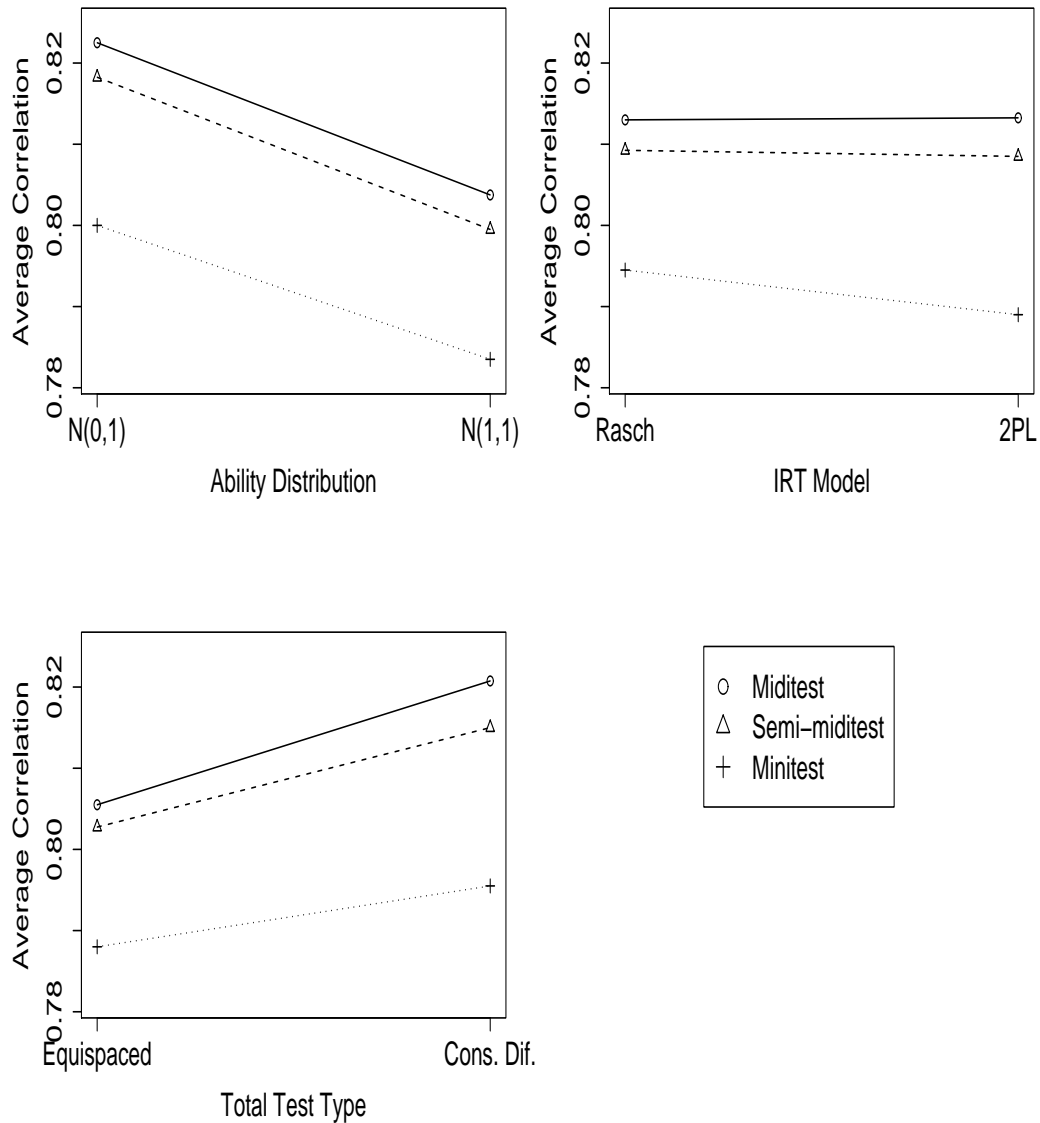


Figure 3. Plots showing the first-order interaction effects involving the anchor tests and the other three factors in the simulation studies for univariate IRT models.

Note. The top left plot shows the average values for the 12 combinations of ability distribution and anchor test, averaging over the other two factors (IRT model and total test type). The top right plot shows the average values for the 12 combinations of IRT model and anchor test. The bottom left plot shows the average values for the 12 combinations of total test type and anchor test. The vertical scale is the same for all the three plots.

4. The average correlation is higher for the constant difficulty total test than for the equispaced total test.
5. The average correlation is higher for the $\mathcal{N}(0, 1)$ case (i.e., when the total test and the anchor test are targeted to the population) than for the $\mathcal{N}(1, 1)$ case.
6. The effect of the Rasch model versus the 2PL model is very small compared to the other factors.

We also examined the correlations for a few other types of anchor tests known to be suboptimal to find out how poorly they perform with respect to correlation. To our surprise, they did not perform as badly as we expected. For example, an anchor test with difficulty parameters equispaced between 0 and 2 never had average correlation lower than that for the minitest by more than 0.02. Though subject to further research, this phenomenon suggests that it may be possible to relax stringent restrictions on anchor test item difficulties without causing serious problems.

2.2 Simulations for Bivariate Tests

The simulations above deal with tests that measure only one skill; in this part, we consider tests that measure two skills. We generate data under a 40-item two-dimensional IRT model (e.g., Reckase, 1997) with item response function

$$P(y_j = 1|\theta_1, \theta_2) = (1 + e^{-(a_{1j}\theta_1 + a_{2j}\theta_2 - b_j)})^{-1}, \quad (\theta_1, \theta_2)' \sim \mathcal{N}_2(\mu_1, \mu_2, 1, 1, \rho), \quad |\rho| \leq 1. \quad (2)$$

This models a test that captures two different, but correlated skills. To simplify, we assume a test in which each item measures only one skill. The first half of the items (20 of them) are assumed to measure the first skill, and the last half are assumed to measure the second skill. Mathematically, ensure this by assuming

$$a_{1j} = \begin{cases} 1 & \text{if } j \leq 20 \\ 0 & \text{otherwise} \end{cases}$$

and $a_{2j} = 1 - a_{1j}$.

Results for three simulation cases, all with $\mu_1 = \mu_2$ equal to 0, but with ρ equal to 0.5, 0.7, and 0.9, are reported here. The difficulty parameters for the total test for all

simulation cases are equispaced between -2 and 2 for both the first 20 items and the last 20 items. The first 10 items in the anchor tests measure the first skill and the last 10 items measure the second skill; that is, the anchors have the same content coverage as the total test. For the minitest, the difficulty parameters are equispaced between -2 and 2 for both the first 10 items and also for the last 10 items. For the miditest, the difficulty parameters are 0 for all items. For the semi-miditest, the difficulty parameters are equispaced between -1 and 1 for both the first 10 and the last 10 items. We also consider two more anchors that do not have the same content coverage as the total test and measure only the first skill (i.e., they have $a_{1j} = 1, a_{2j} = 0$ for all j). The first of these has equispaced difficulties between -2 and 2, while the second has all difficulties equal to 0. We call these the content nonrepresentative (CNR) minitest and CNR miditest, respectively.

Table 2.
The Average (and SD, Within Parentheses) of the Correlation Coefficients for Data Generated From a Bivariate IRT Model

Anchor	$\rho = 0.5$	$\rho = 0.7$	$\rho = 0.9$
	Correlation coefficient	Correlation coefficient	Correlation coefficient
Mini-test	0.737 (0.014)	0.759 (0.014)	0.777 (0.013)
Semi-miditest	0.757 (0.012)	0.777 (0.011)	0.794 (0.011)
Midi-test	0.762 (0.012)	0.782 (0.012)	0.798 (0.011)
CNR minitest	0.660 (0.016)	0.713 (0.015)	0.764 (0.013)
CNR miditest	0.673 (0.016)	0.727 (0.013)	0.780 (0.011)

The average correlation coefficients (along with SDs) for each simulation case from 100 replications each are shown in Table 2. Figure 4 shows the correlation coefficients graphically.

The comparative performance of the different anchor tests is very similar to that

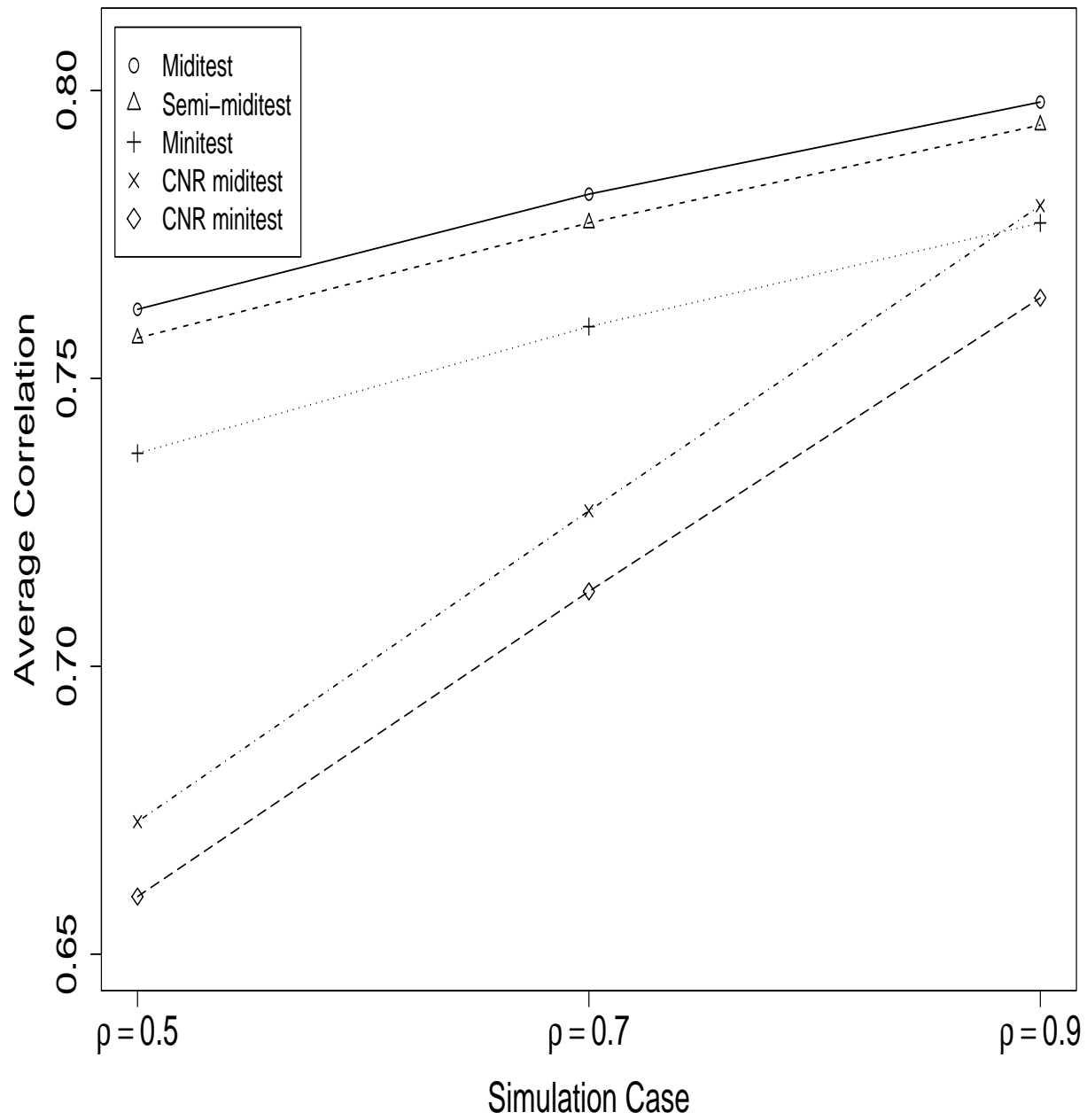


Figure 4. Plot of the average correlation coefficients for the multivariate IRT model under different scenarios.

in the univariate simulations. The average correlations are highest for the miditest again, closely followed by those for the semi-miditest. And, once again, the minitest is the worst performer with respect to correlation. The values of the correlation coefficient are comparatively lower for the CNR anchor tests, which supports the findings of Klein and Jarjoura (1985) regarding the importance of content representativeness of anchor tests. Also, the correlation coefficient between the scores in the total test and the anchor test increases as ρ , the correlation between the two skills, increases. One interesting finding is that, for high correlation, the anchor-test to total-test correlation for the CNR miditest is comparable to that for the minitest. A high correlation in this instance might be $\rho = 0.9$, which is not rare in practice; for example, the correlations between components of the proficiency vector in several assessments in National Assessment of Educational Progress are around 0.9 (see, e.g., von Davier & Sinharay, in press).

2.3 Discussion

The above simulations considered external anchors only. We found that the results are the same for internal anchors as well (i.e., the miditest has the highest correlations with the total test; results not shown), as well as for other values of sample size, total test length, and anchor test length. These findings demonstrate that the miditest and the semi-miditest (which have statistical characteristics that differ from those of the equispaced total test) are the winners from the viewpoint of anchor-test to total-test correlation.

2.4 Further Simulations

The question remains as to what happens if the miditest and semi-miditest (which, the above results suggest, should ideally be centered at the mean of the proficiency distribution) are not centered properly. Do they still perform better than the minitest? To answer that question and to study the interactions among (a) the mean item difficulty of the total test, (b) the mean item difficulty of the anchor test, (c) the variance of the item difficulties of the anchor test, and (d) the mean of the proficiency distribution, we performed another set of simulations. A primary goal in these simulations was to examine how the anchor-test to total-test correlation varies as the item difficulty distribution of

the anchor test differs from that of the total test (note that the common belief is that they should not differ, which happens for the minitest). Data were generated for 2,000 examinees from the Rasch model. The total test had 60 items, while the anchor test had 30 items. The generating proficiency distribution is always a normal distribution with SD 1. The difficulty parameters for the total test and the anchor test were generated from a normal distribution, the mean of the distribution being 0 for the total test. The following four factors were controlled in these simulations:

- The SD of the difficulty parameters of the total test: Three values (0.1, 0.5, and 1) were used.
- The SD of the difficulty parameters of the anchor test: Three values (0.1, 0.5, and 1) were used. The value 0.1 corresponds to a miditest, the value 0.5 corresponds to a semi-miditest, and the value 1 corresponds to a minitest.
- The mean of the difficulty parameters of the anchor test: Three values (0, 0.25, and 0.5) were used. A mean of 0.25 denotes some departure of the anchor from the mean ability, while the mean of 0.5 denotes substantial departure.
- The mean of the generating proficiency distribution: Five values (-0.5, -0.25, 0, 0.25, and 0.5) were used.

For each of the $3 \times 3 \times 3 \times 5 = 135$ simulation cases, we generated 1,000 replications of total test and anchor test (the generating difficulty parameters were allowed to vary over the replications), and we computed the anchor-test to total-test correlation. Figure 5 shows the average correlation coefficients (averaged over the 1,000 replications) for the 27 simulation cases when mean of the proficiency distribution is 0.

The variation between the average correlations is quite low (the range is 0.860 to 0.878); however, the figure shows several patterns:

- The average correlation increases as the SD of the difficulty parameters of the total test decreases. This was found in the earlier simulations as well.
- The average correlation increases as the mean of the difficulty parameters of the anchor

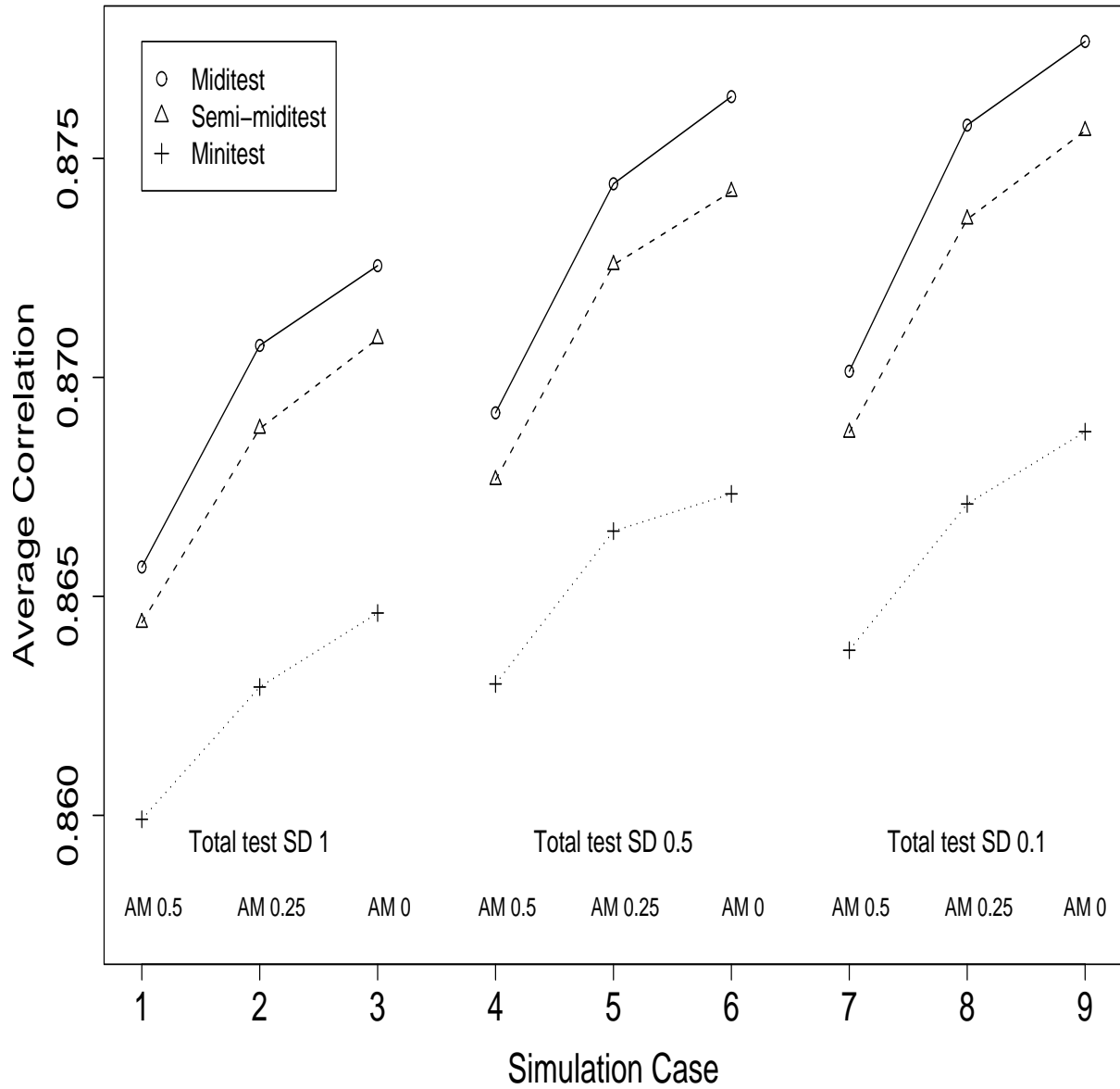


Figure 5. Plot of the average correlation coefficients for different values of the SD of the difficulty parameters of the total test and mean and the SD of the difficulty parameters of the anchor test.

Note. The mean of the proficiency distribution and mean difficulty of the total test are 0. The mean difficulty of the anchor test is denoted as *AM* in the legend of the plot.

test approaches 0 (i.e., as the anchor test is more targeted to the examinee distribution and the total test).

- Most importantly, the miditest and the semi-miditest have considerably higher average correlation than the minitest in all the simulation cases.
- Interestingly, a targeted minitest (i.e., with a mean of 0) always has less average correlation than an off-target (mean of 0.5 or 0.25) miditest or semi-miditest.

Figure 6 shows the average correlation coefficients for the 27 simulation cases where mean difficulty of the anchor test is 0, and the mean of the proficiency distribution is -0.5, -0.25, or 0.

This figure, again demonstrating little variation in the average correlations, shows that that the average correlation increases as the

- SD of difficulty of anchor test items decreases (i.e., as the anchor test approaches a miditest)
- SD of difficulty of total test items decreases
- mean of the proficiency distribution gets closer to the mean of difficulty (0) of total and anchor test items

Figure 7 shows the average correlation coefficients for the 27 simulation cases where SD of the difficulty parameters of the total test is 1, and the mean of the proficiency distribution is -0.5, 0, or 0.5.

This figure, in spite of showing some interactions (e.g., when proficiency mean increases from 0 to 0.5, average correlations for anchors with mean 0.5 increase, but average correlations for anchors with mean 0 or 0.25 decrease), shows that the average correlation is higher for the miditest and the semi-miditest than for the minitest.

Figures 6 and 7 show, among other things, what happens for anchor tests that are not targeted to the population or to the total test. While the correlations are lower for the noncentered anchor tests than the centered anchor test, the miditest or the semi-miditest still improve upon the minitest with respect to correlation.

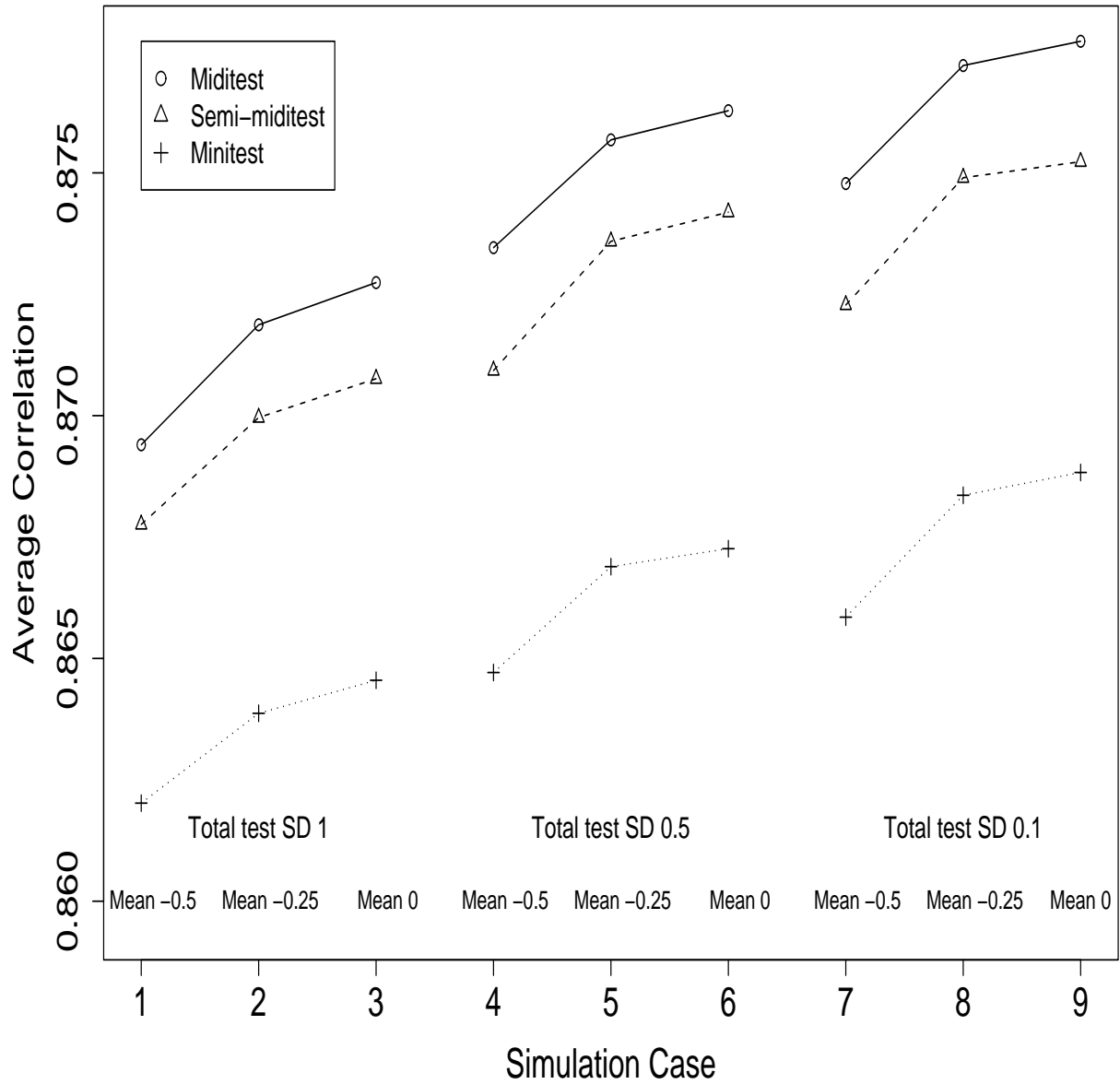


Figure 6. Plot of the average correlation coefficients for different values of SD of the difficulty parameters of the total test and anchor test and mean of the proficiency distribution.

Note. The mean difficulty of the total test and the anchor test are 0. The *means* in the legend of the plot denote the mean of the proficiency distribution.

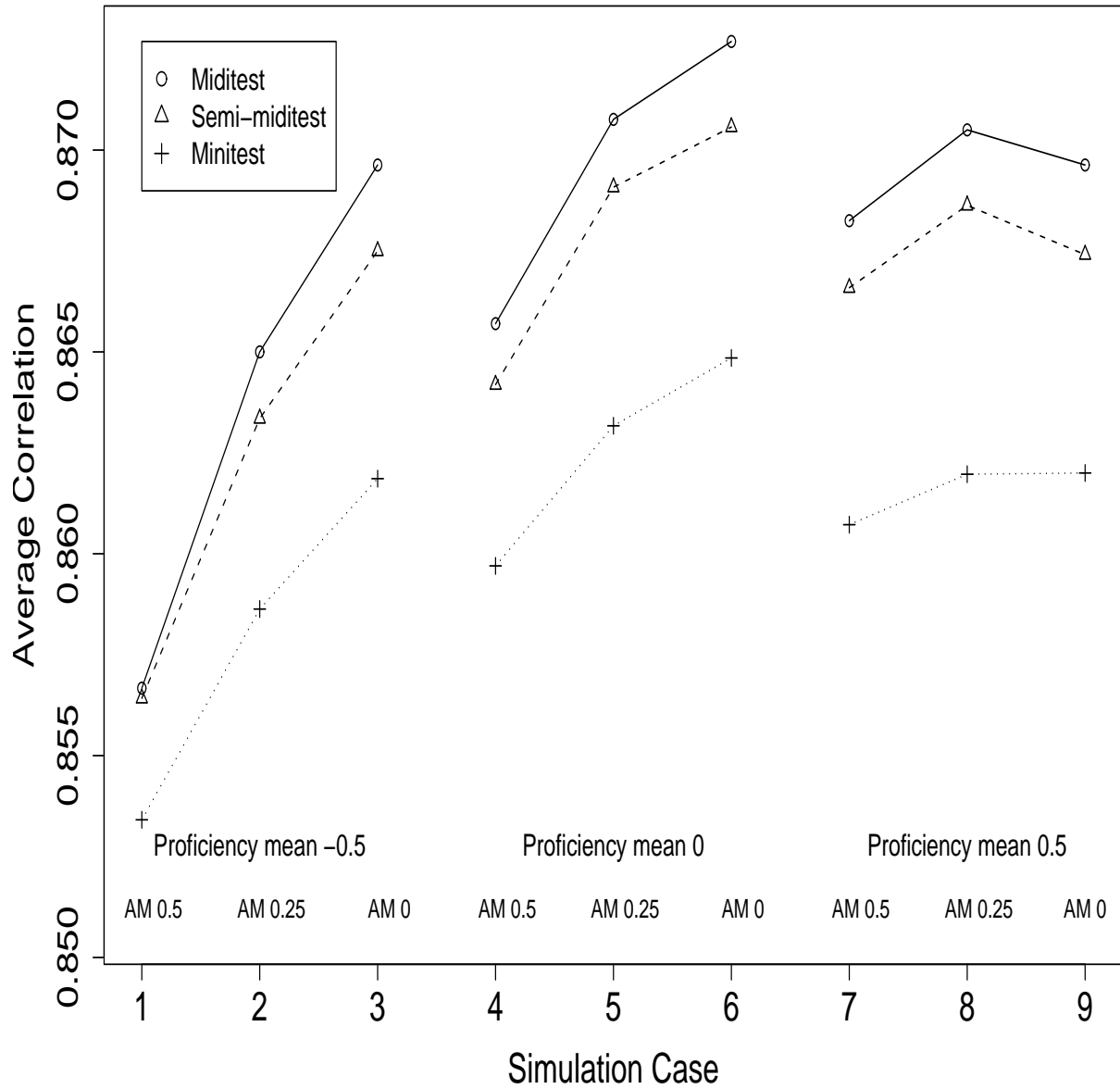


Figure 7. Plot of the average correlation coefficients for different values of the mean of the proficiency distribution and mean and SD of the difficulty parameters of the anchor test.

Note. The mean and SD of the difficulty parameters of the total test are 0 and 1, respectively. The mean difficulty of the anchor test is denoted as it AM in the legend of the plot.

Thus, although a miditest or a semi-miditest centered at the same point as the total test and the proficiency distribution has the highest anchor-test to total-test correlation, the above simulations suggest that, even if they are a bit off target, the correlation will still be quite high—and higher than that of the minitest.

3. Real Data Example

We also examined data from an operational basic skills test. The test has 6,489 examinees and 110 multiple-choice items that are scored in 0 or 1. The items are on six content areas, such as mathematics and science. The operationally used internal anchor test (which was created to be a minitest) with 34 items, which are approximately equally distributed among the six content areas, acts as a minitest. To create a miditest, we first sorted all the items in the total test within each content area according to difficulty (proportion correct). The data set did not have too many items with exactly medium difficulty to form a true miditest. Therefore, we opted for a semi-miditest; to obtain that, we chose the same number of items within each content area as in the minitest, and, for each content area, we chose the items that were middle-most according to the values of proportion correct. The average proportion correct for the total test, minitest, and the semi-miditest are 0.72, 0.73, and 0.75, respectively, so that the center of the minitest is closer to that of the total test than that of the semi-miditest. The reliabilities of the total test, the minitest, and the semi-miditest are 0.86, 0.66, and 0.69, respectively.

The correlation coefficient between the total test scores and the minitest scores is 0.875, while that between the total test scores and the semi-miditest scores is 0.893. Thus, as in the simulations, the semi-miditest outperforms the minitest with respect to correlation. It is also interesting to note that if we form an internal anchor test ignoring the content areas and just choosing the 34 items with the middle-most values of proportion correct (this test has average proportion correct of 0.73), the correlation coefficient between its scores and the total test scores is 0.904, which supports the finding from the bivariate simulation that a CNR miditest can have a higher correlation than a minitest.

Figure 8 shows the distribution of the raw scores of the minitest and the semi-miditest. It is clear that the semi-miditest is slightly easier than the minitest.

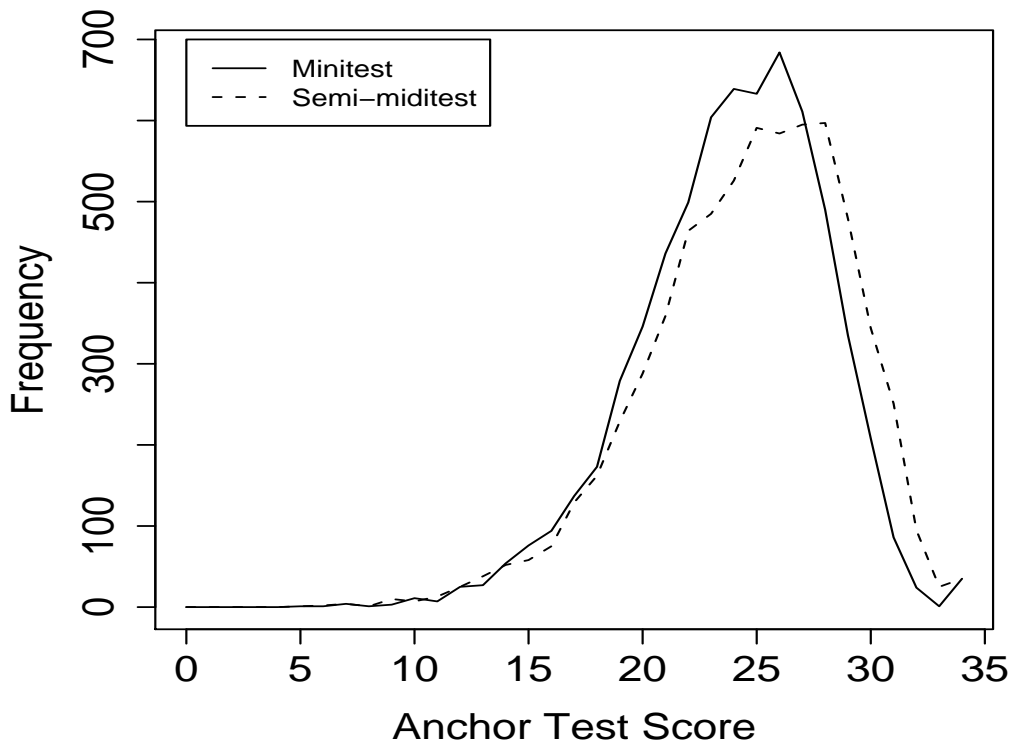


Figure 8. Plot of the distribution of raw scores of the minitest and the semi-miditest for the basic skills test.

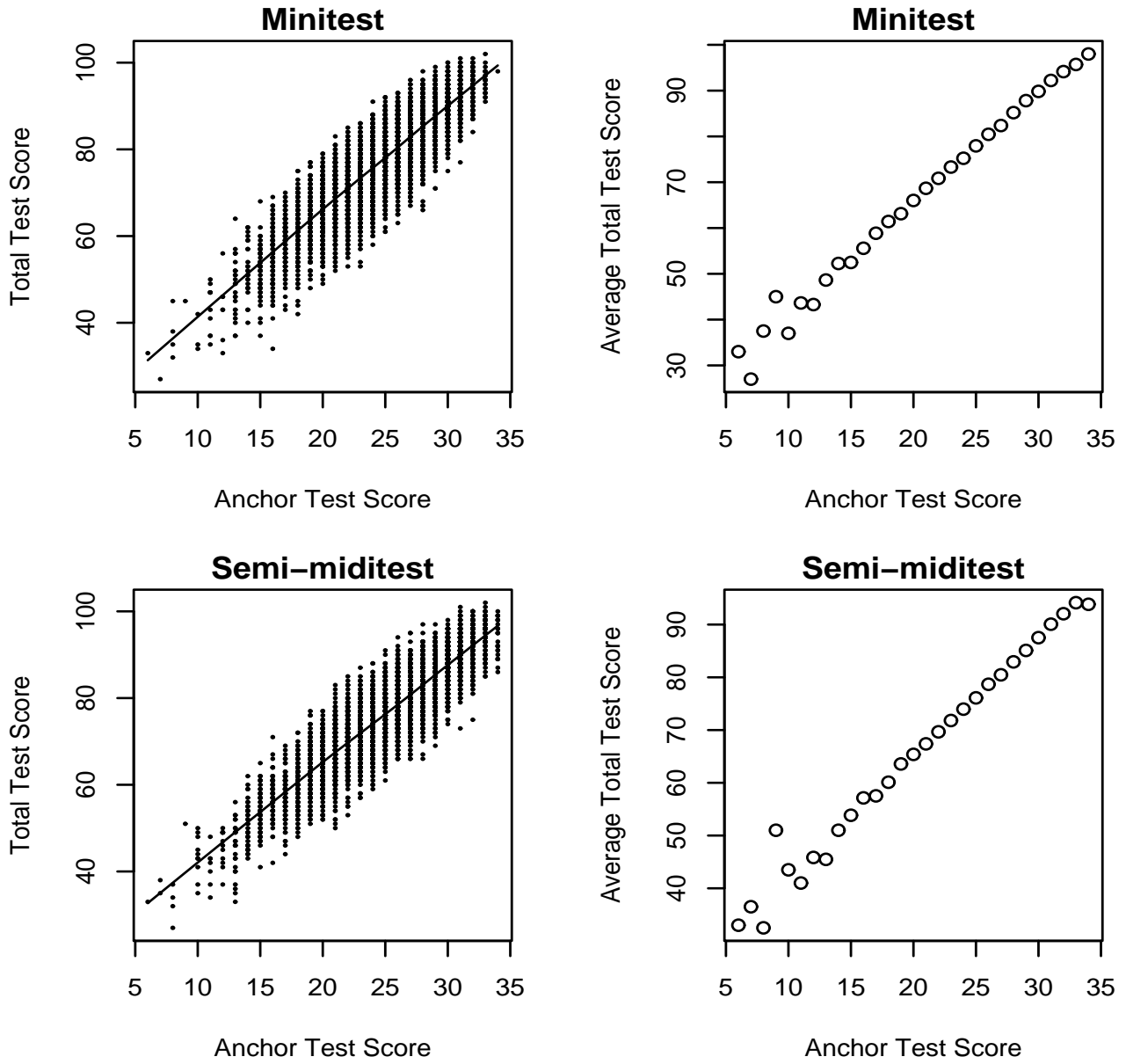


Figure 9. Plot of the raw scores of the total test vs. the minitest (top row) and semi-miditest (bottom row) for the basic skills test.

Note. In both rows, the left panel shows a scatter-plot of the raw scores in the total test versus that in the anchor test, and the right panel shows the average total test score for any anchor test score.

Figure 9 shows the raw scores of the minitest versus those in the total test (top left panel), and the raw scores of the semi-miditest versus those in the total test (bottom left panel).

To further study the relationship between the raw scores, the LOWESS regression curve (Cleveland, 1979), a nonparametric curve that best fits the data locally, is also shown on the left panels. The right panels plot the values of the average total test score for any given anchor test score, which is like the regression of the total test score on the anchor test score.

Figures 8 and 9 show hardly any difference between the minitest and the semi-miditest regarding the relation of their scores and the total test scores and do not point to any obvious limitations of the semi-miditest. Figure 9 shows that the relationship between the total test scores and the semi-miditest scores is almost linear. Hence, the correlation coefficient between the two sets of scores is an appropriate summarization of their relationship, and, because anchor-test equating exploits the relationship between the total test scores and anchor test scores, the figure justifies examining correlation coefficient in this paper and makes us confident that the higher correlation for the semi-miditest will lead to better equating.

Figure 10 demonstrates in another way the superiority of the semi-miditest scores in linearly predicting the total test scores, by showing a *binned residual* plot. To create this plot, we run linear regression of total test scores on the (a) minitest scores, and (b) semi-miditest scores and then compute the residuals. For each regression, we compute the nine deciles of the anchor test scores that define 10 intervals and then compute the average absolute value of the residuals for each interval. Figure 10 plots the average absolute residual against the average anchor test score for each of the ten intervals. Though the average anchor test scores are slightly different for the two anchor tests in each interval, it is clear that the average absolute residuals are mostly smaller for the semi-miditest than for the minitest, the difference being significant for the left-most interval. Note that lower absolute residuals in linear regression leads to higher correlation and vice-versa. The same pattern as in Figure 10 is observed if we create a plot (not shown) using the residuals from the LOWESS regression.

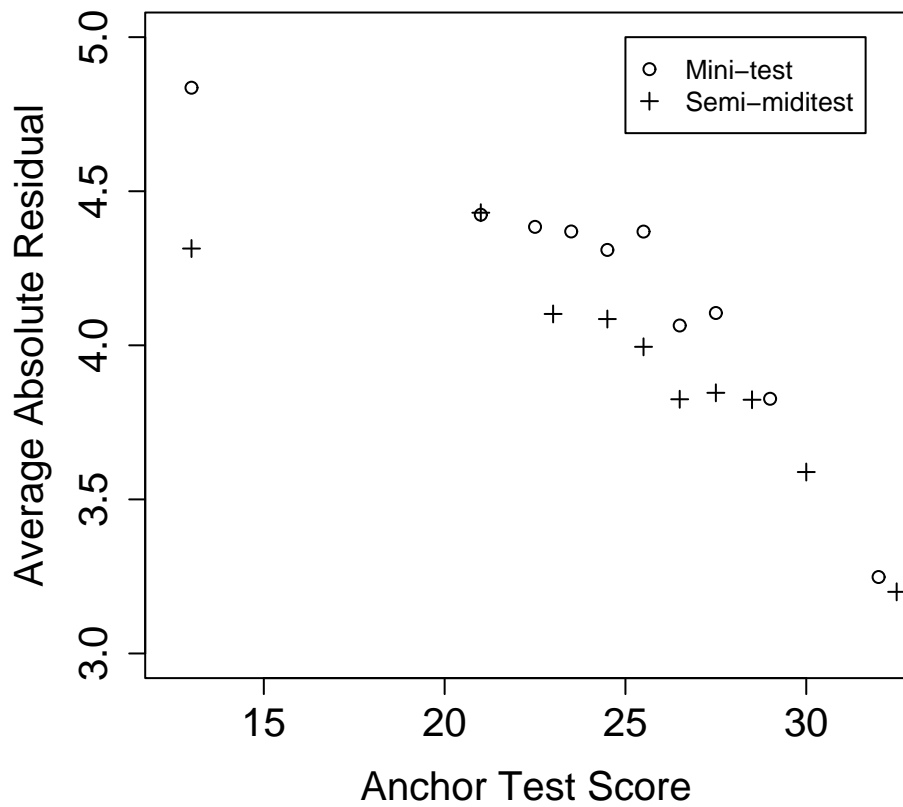


Figure 10. Binned residual plot for linear regression of total test scores on the (a) minitest scores and (b) semi-miditest scores for the basic skills test.

4. Discussion and Future Work

This paper is a preliminary examination of whether an anchor test has to be statistically representative of the total test. Through a series of simulated data (generated from unidimensional and multidimensional IRT models) and real data examples, we show that the traditionally recommended minitest is not the optimum anchor test so far as the correlation between an anchor test and the total test is concerned. The anchor-test to total-test correlation is shown to be consistently lower for the minitest than for an anchor test whose spread of item difficulties is less than that of a total test.

Our results, together with the common knowledge that *higher correlation leads to better equating* (e.g., Angoff, 1971; Petersen et al., 1989; von Davier et al., 2004), imply that there is a need for more thorough studies using anchor tests that relax the assumption on the spread of item difficulties, especially in the context of actual equating situations. As discussed earlier, the SEE will be lower for such tests (e.g., a *miditest* or a *semi-miditest*) than for the minitest for the random groups with anchor test design (Angoff, 1971; Lord, 1950). For the other equating methods and designs such as the Tucker/Levine/chain equating methods for a NEAT design, it is not straightforward to show a simple relationship like Equation 1 between the SEE and the correlation coefficient—these are areas that require further research. It will also be interesting to examine the performance of these newly suggested anchor tests in nonlinear equating (e.g., equipercentile equating), where the role of the correlation coefficient is less clear. However, the plot of the scores of total test versus those of the semi-miditest (Figure 9) for an operational test data, where the relationship is almost linear (the same phenomenon happened for the simulated data also—plots not shown here), suggests that the semi-miditest will probably perform respectably even in nonlinear equating.

Our findings suggest that to compare the performance of a minitest and the newly suggested anchor tests in actual equating conditions, anchor tests with varying degrees of spread of item difficulties should be employed and their performances compared in order to strike a balance between the statistical qualities of an anchor and its operational convenience. For example, a perfect miditest (i.e., an anchor test with very small spread of item difficulties for each content area) has the highest anchor-test to total-test correlation;

but such a test will be difficult to obtain operationally. However, a semi-miditest (an anchor test with moderately high spread of item difficulty for each content area) will be easy to obtain operationally and has an anchor-test to total-test correlation quite close to that for the miditest. Though our findings show that a semi-miditest targeted at the same point as the total test and the ability distribution is the optimum, they also show that semi-miditests that are slightly off target do not perform too poorly, and it might be worthwhile to study such anchor tests (those, if found good, may provide considerable flexibility operationally, as the administrators then do not have to worry about the anchor test being exactly on target). Another important issue is the choice of the performance criteria in such equating studies. Standard error of equating will be a natural criterion. However, to study bias of equating, another natural criterion, one needs to know the truth, which is very difficult to obtain. A study design as in von Davier, Holland, and Livingston (2005) can be employed in this respect; in that study, a long test was administered to two groups of examinees, P and Q, and was partitioned into an old form X, new form Y, and an anchor test A. Then, pretending that the population P was not administered the test Y, and the population Q was not administered the test X, the scores in X were equated to the scores in Y using the anchor A (several equating methods were employed). The equating function of X to Y using a single group design (as both P and Q actually took both X and Y) was used as a criterion/truth in the comparison. However, studies like von Davier et al. are not straightforward enough to perform operationally.

If the future equating studies suggested above show that anchor tests that relax the assumption on the spread of item difficulties perform better than a minitest in actual equating situations, it will have a significant effect on test development. Because medium difficulty items are more easily available than items with extreme difficulty, this work will then provide test developers with more flexibility while constructing anchor tests. On the other hand, if these newly suggested anchor tests perform worse than a minitest in equating, it will suggest that the recommendation that higher correlations lead to better equating should be taken with caution; such a finding will probably improve our understanding of when and how an anchor test with high correlation may not perform well in equating. Thus, this work is likely to be useful regardless of how it turns out.

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.
- Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of American Statistical Association*, *74*, 829-836.
- 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.
- Cronbach, L. J., & Warrington, W. G. (1952). Efficacy of multiple choice tests as a function of spread of item difficulties. *Psychometrika*, *17*, 127-147.
- 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, Quebec.
- von Davier, M., & Sinharay, S. (in press). An importance sampling EM algorithm for latent regression models. *Journal of Educational and Behavioral Statistics*.
- 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.
- Gulliksen, H. (1945). The relation of item difficulty and inter-item correlation to test variance and reliability. *Psychometrika*, *10*(2), 79-91.
- 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.
- Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: ETS.
- Lord, F. M. (1950). *Notes on comparable scales for test scores* (ETS RB-50-48). Princeton, NJ: ETS.
- Lord, F. M. (1952). The relationship of the reliability of multiple-choice test to the distribution of item difficulties. *Psychometrika*, 18, 181-194.
- Michaelides, M. P., & Haertel, E. H. (2004). *Sampling of common items: An unrecognized source of error in test equating*. Submitted for publication.
- Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating method (pp. 71-135), In P. W. Holland & D. B. Rubin (Eds.). *Test equating*, 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.
- Richardson, M. W. (1936). The relationship between difficulty and the differential validity of a test. *Psychometrika*, 1, 33-49.