Research Report

An Alternative Data Collection Design for Equating With Very Small Samples

Gautam Puhan

Tim Moses

Mary Grant

Fred McHale

March 2008

ETS RR-08-11

Listening. Learning. Leading.®

An Alternative Data Collection Design for Equating With Very Small Samples

Gautam Puhan, Tim Moses, Mary Grant, and Fred McHale
ETS, Princeton, NJ

March 2008

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

Abstract

A single group (SG) equating design with nearly equivalent test forms (SiGNET) design was developed by Grant (2006) to equate small volume tests. The basis of this design is that examinees take two largely overlapping test forms within a single administration. The scored items for the operational form are divided into mini-tests called testlets. An additional testlet is created but not scored for the first form. If the scored testlets are Testlets 1–6 and the unscored testlet is Testlet 7, then the first form is composed of Testlets 1–6, the second form is composed of Testlets 2–7, and Testlets 2–6 are common to both test forms. They are administered as a single *administered form,* and when a sufficient number of examinees have taken the administered form for an SG equating, the second form (Testlets 2–7) is equated to the first form (Testlets 1–6) using SG equating. As evident, there are at least two merits of the SiGNET design over the nonequivalent groups with anchor test (NEAT) design. First, it facilitates the use of an SG equating design, which has the least random equating error, and second, it allows for the accumulation of sufficient data to equate the second form. Since the examinees scores are based on only the first form (i.e., the operational form), the two forms can be administered until sufficient data are collected to equate the second form. This study compared equatings under the SiGNET and NEAT designs and found reduced bias and error for the SiGNET design in very small sample size situations (e.g., N = 10 or 15). Implications for practice using the SiGNET design are also discussed.

Key words: Small sample, equating, SiGNET, error, bias

Acknowledgments

The authors would like to thank Neil Dorans, Sooyeon Kim, and Dan Eignor for their helpful comments on an earlier draft of the paper and Kim Fryer for editorial assistance.

The phrase *single group nearly equivalent test* or SiGNET was suggested by Neil Dorans, and we also would like to thank him for that.

Table of Contents

List of Tables

List of Figures

Introduction

In order to prevent overexposure and avoid undesirable practice effects, testing organizations use multiple, parallel forms of a single test in different test administrations. Although every effort is made to create parallel forms, some forms turn out to be relatively easier or harder than others. This is unfair for test takers who happened to take the harder test. Test equating is used to adjust for differences in difficulty across different forms of a test, which in turn allows for score comparisons across different groups of test takers regardless of the test forms they were administered or time periods when they took the test.

Like many statistical procedures, test equating is affected by sampling error. If the sample used to derive the equating relationship between the new and old forms is large and representative of the population, then the equating would likely be accurate (i.e., have less sampling error and bias). However, if the sample used to conduct the equating is very small, then the equating could be inaccurate (Livingston, 1993). Testing programs frequently encounter less than optimal sample sizes. This number is often less than 50 and in some cases even less than 10. Consequently, sample size poses a problem when a new test form is introduced and has to be equated to an old form already on scale (i.e., equating error is substantially large).

Small samples may occur for several reasons. One possible reason is that some tests are very specialized (e.g., tests measuring knowledge about Latin, theatre, etc.), and therefore fewer examinees take these tests compared to other tests. Another possible reason is that testing programs often administer tests several times a year (known as test administrations). Although the number of examinees taking a particular test may be large for a whole testing year, the number of test takers in each test administration may still be quite low. Regardless of the reasons for small sample sizes, equating is needed when new forms are introduced and scores have to be reported to test takers on time.

The purpose of this study is to describe and evaluate a single group pre-equating design using testlets, which may prove useful when equating with small samples. This design was described extensively by Grant (2006), who presented the details of this design and showed its usefulness in equating under small sample size conditions. This study will briefly describe the small sample equating design and then present an application of this design using real test data under hypothetical testing conditions.

The basic idea of this small sample equating design is that two largely overlapping test forms are administered to a group of examinees in a single administration. The first form constitutes the operational form (i.e., examinees are scored on this form). The second form is administered along with the first form to gather pretest data for the second form, which can be used later to equate the second form to the first form using a single group (SG) equating design. This procedure is repeated to equate future new forms. Because this design involves equating two largely overlapping test forms using the SG equating design, this design is referred to as the *single group nearly identical tests* design, or SiGNET design, throughout this paper.

Previous Studies on Small Sample Equating

What are recommended sample sizes for equating, and how does one equate when the sample size is unavoidably small? Several previous studies have tried to answer these important questions. For example, to address the first question, Kolen (1985) examined samples of 100 and 250 and found standard errors of equating, when derived without the normality assumption, to be sufficiently accurate with sample sizes of 250. Similarly, Skaggs (2005) studied equating using samples ranging from 25 to 200 in an equivalent groups design and concluded that for samples as small as 25, no equating was likely to be preferable, but for samples ranging from 50 to 75, equating was preferable to no equating.

Addressing the second question, Livingston (1993) focused on the effectiveness of log-linear presmoothing with samples of 25, 50, 100, and 200 examinees per form, using the nonequivalent groups with anchor test (NEAT) design. Livingston found that presmoothing significantly reduced equating error, particularly for the smallest samples. Equating error using presmoothing was about as effective as unsmoothed equating using samples twice as large. Similarly, Hanson, Zeng, and Colton (1994) compared linear and identity equating (no equating) with unsmoothed, presmoothed, and postsmoothed equipercentile equating for five ACT assessment tests. They found that smoothing significantly improved equipercentile equating with small samples, although there was no clearly preferred pre- or postsmoothing method.

Although presmoothing has been shown to improve equating in small samples, it is still a matter of debate whether it produces accurate results in very small samples. According to Holland, Dorans, and Petersen (2007) and Petersen (2007), smoothing helps in equating for moderate sample sizes but may not be of much help for very small samples. This is especially true when it is unclear how well the small sample represents the intended population. It is also

worth noting that if the samples are very small, then log-linear smoothing, a nonlinear process, may not correct for the sampling bias due to inadequate sampling, and this can counteract the gain due to reduction in the standard error of equating (Kim, von Davier, & Haberman, 2006).

In a more recent study, Kim et al. (2006) proposed using an average (which they referred to as the synthetic linking function) between the identity function and an estimated equating function based on a small sample. Although the identity seemed preferable for very small samples (i.e., $N < 25$), the synthetic function seemed preferable for moderately small samples such as 50 and 100. The authors also suggested that the synthetic linking function may be a good choice when the test specifications are well defined and the test forms are close to parallel in content and difficulty.

The current study fits better with research to address the second question—how to equate when the sample size is unavoidably small. However, this study differs from some of the previous studies that addressed this question in one important way. While previous studies tried to incorporate modifications to existing equating methods to improve equating under small sample conditions (e.g., presmoothing before equating, averaging two equating functions), the SiGNET design proposes a change to the data collection design such that all scored items in a new form are present on previous forms as either scored or nonscored items. This, in turn, facilitates the use of an SG equating design, which has been shown to have much less error (see Thorndike, 1982 and Kolen & Brennan, 2004) than other equating designs such as the NEAT design. Since proper data collection is regarded as the most important aspect of any equating (see Holland et al., 2007 Holland & Dorans, 2006), the proposed SiGNET equating design, whereby data conducive to improved equatings can be collected, makes this approach an advantageous one to employ.

Description of the Single Group Nearly Identical Tests (SiGNET) Design

The basis of the SiGNET design for small samples is such that a group of examinees takes two largely overlapping forms of a test as a single combined form (during a single administration). The scored items for the operational test form are divided into several testlets of an equal number of items (see Figure 1 for an example). Each testlet[1] matches the total test specifications as closely as possible. An additional testlet is created, also matching the specifications as closely as possible. This additional testlet will not be scored and is only used for data collection. If the scored testlets are designated as Testlets 1–6 and the additional unscored

testlet is designated as Pretest Testlet 7, then the first form (Form 1) is composed of Testlets 1–6 and the second form (Form 2) is composed of Testlets 2–7. They are administered as a single *administered test form* in the first administration, where the examinees do not know which items belong to each testlet. When a sufficient number of examinees have taken the administered form for a single group equating, the second form (Testlets 2–7) is equated to the first form (Testlets 1–6) using a single group equating method. Similarly, when the second form (Testlets 2–7) is administered, an additional pretest Testlet 8 is created and administered along with the second form as a single administered form. When a sufficient number of examinees has taken the administered form for a single group equating, the third form (Testlets 3–8) is equated to the second form (Testlets 2–7) using a single group equating method.

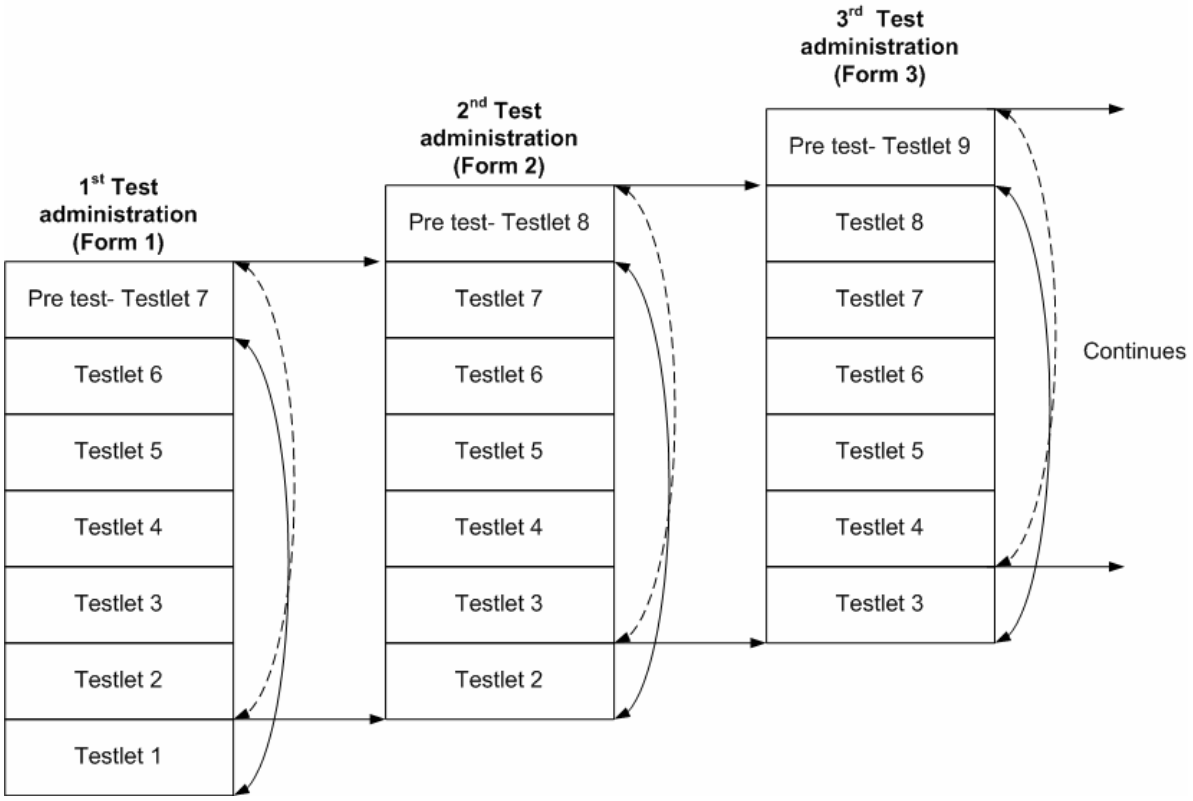

Figure 1. **Flowchart showing the small sample single group nearly identical tests (SiGNET) data collection and equating design.**

Note that Form 2 was really a pseudo-form in the first test administration and became an operational form in the second test administration. In reality, it may take several administrations before pseudo-Form 2 becomes operational, because it may take several administrations to

accumulate sufficient data to conduct a single group equating that will be applied to the second form. Similarly, Form 3 was a pseudo-form in the second test administration and became an operational form in the third test administration. In Figure 1, the bold and dashed curved lines with the two-sided arrows represent the operational and pseudo-forms, respectively, for each administration.

Although under the SiGNET equating design there is a slow change of test items from one form to another, the process eventually reaches a stage where all the items in the first form are replaced by new items in a future form (e.g., Form 7 or 8). Some may argue that having a large overlap between the new and old forms and delivering the same form in repeated administrations to accumulate enough data for equating under the SiGNET design increases the risk of exposure. In reality, however, since this design is proposed for very small volume tests with test takers often testing in different parts the country, the risk of overexposure may be minimal compared to that for high volume tests. In the case of high volume tests, even though new forms are introduced more frequently, there is still some overlap of items (i.e., anchor items) which are exposed to much larger testing groups. As a separate note, testing programs with very small volumes typically administer one or two forms for repeated administrations because of the inability to equate a new form if it is introduced. Therefore the SiGNET design does not necessarily increase overexposure in those cases. In fact, it allows for a slow change, as opposed to allowing relatively little or no change over a longer period of time.

As evident, a strong feature of the SiGNET equating model is that it allows new forms to be equated using an equating method that has the smallest random equating error (i.e., the SG equating method). As an added benefit, this model allows the testing practitioner to accumulate data over different administrations before conducting an SG equating of the new form, which in turn may help reduce sampling error and bias. Since the examinees' scores are based on only the first form (i.e., the operational form), the two forms can administered repeatedly until sufficient data are collected to equate the second form. Under the NEAT design, although the old form sample can be accumulated to get a larger volume, the new form sample will still be very small, which can lead to large equating error. Considering these advantages (i.e., equating using an SG equating method and the opportunity to accumulate samples over administrations to increase sample size) and limitations (i.e., potential risk of overexposure because of considerable overlap between new and old forms and delivering the same form for repeated administrations to

accumulate enough data to equate), the SiGNET model may be viewed as a compromise between the NEAT model, where new forms with less overlap are introduced more frequently, and a situation where the same form is repeated for many administrations.

Purpose of the Study

The purpose of this study is to compare the accuracy of equating derived using the SiGNET design with that derived using the NEAT design in small sample size situations. If the SiGNET design were not available, then new forms would normally be equated to the old form using common items (i.e., the NEAT design). Therefore comparison of the SiGNET design to the NEAT design seems reasonable.

In the example shown in Figure 1, one can derive two equating conversions for Form 3. The first conversion can be based on the SiGNET equating design, whereby Form 3 is linked to the base scale (Form 1) via Form 2. It involves a chain of equatings where the conversion for Form 3 is based on an SG equating (i.e., pseudo-Form 3 equated to Form 2) using the examinees who took Form 2, and the conversion for Form 2 is also based on an SG equating (i.e., pseudo-Form 2 equated to Form 1) using the examinees who took Form 1.

The second conversion can be based on a NEAT equating. Since Form 3 and Form 1 share common items (i.e., Testlets 3–6), Form 3 can be directly equated to Form 1 using these common items. Note that in reality the sample size for Form 3 under the NEAT design would be small, whereas under the SiGNET design, data could be accumulated before the form is operationally scored and therefore pre-equated using a larger accumulated sample. Although the larger sample size and the SG equatings may improve the equating under the SiGNET design, the introduction of additional equating chains may result in higher overall error and inaccuracy. Therefore, the purpose of this study is to compare the equating conversions derived via these two approaches (i.e., SiGNET versus NEAT).

Method

Test Data and Procedure

The data for the present study was taken from the responses of 23,000 examinees on one form of a 120-item basic skills test. This test was composed of four content areas: reading, mathematics, social studies, and science, with each content area contributing to 25% of the total test (i.e., 30 out of 120 items). These 120 items were used to create five smaller test forms with

almost 83% overlap between successive forms (i.e., between the new and the most recent old form). Since these five smaller forms were essentially created from one test form for which responses from 23,000 examinees were available, it resulted in a hypothetical testing condition where all the five forms had responses from all 23,000 examinees. This hypothetical testing condition facilitated the creation of a strong criterion equating (which is explained in a subsequent section), to which the small sample equatings were compared.

Creating a hypothetical testing condition to compare the single group nearly identical test (SiGNET) and nonequivalent groups with anchor test (NEAT) designs. The hypothetical testing condition was created as follows. From the 120-item test, 10 testlets comprising 12 items each were created. Since the testlets are supposed to be mini-versions of the total test, items for each testlet were selected such that the content areas (i.e., reading, mathematics, social studies, and science) were represented in the testlets in the same proportion as they were in the total test. After the item selection procedure was completed, each testlet was composed of 12 items, with 3 items belonging to each of the reading, mathematics, social studies, and science content areas. Therefore, similar to the total test, each content area represented 25% of the whole testlet. The 20 testlets had a mean difficulty range of 0.05 (min p+ = 0.63 and max p+ = 0.68), which is close to the mean difficulty of the total test (mean p+ = 0.65). The mean scores for the 10 testlets ranged from 7.56 to 8.21, and the *SD*s ranged from 1.84 to 2.14.

From the 10 testlets that were created, 6 testlets (Testlets 1–6) were combined to create Form 1 (see Figure 2). An additional testlet (Pretest Testlet 7) was added to Form 1. Hence Form 1 was composed of Testlets 1–6, and Form 2 (i.e., the pseudo-form) was composed of Testlets 2–6 and Pretest Testlet 7. Both the forms were administered as a single combined form in several testing administrations, and once enough data were accumulated, an SG equating of Form 1 and pseudo-Form 2 was conducted to put Form 2 on scale. Note that the equating was done using the sample(s) that took both the forms, and the resulting conversion was applied to Form 2 (similar to pre-equating Form 2 before it is administered as an operational form). Also note that it was planned beforehand to drop Testlet 1 from Form 1 to make the transition to Form 2. This process was repeated for equating Form 3. After enough data were accumulated on Form 2 (Testlets 2–7) and pseudo-Form 3 (Testlets 3–7 and pretest Testlet 8), an SG equating of Form 2 and pseudo-Form 3 was conducted to put Form 3 on scale. As before, the equating was done using only the sample(s) that took both the forms, and the resulting conversion was applied to Form 3. Also it

was planned beforehand to drop Testlet 2 from Form 2 to make the transition to Form 3. This process was repeated to put Forms 4 and 5 on scale.

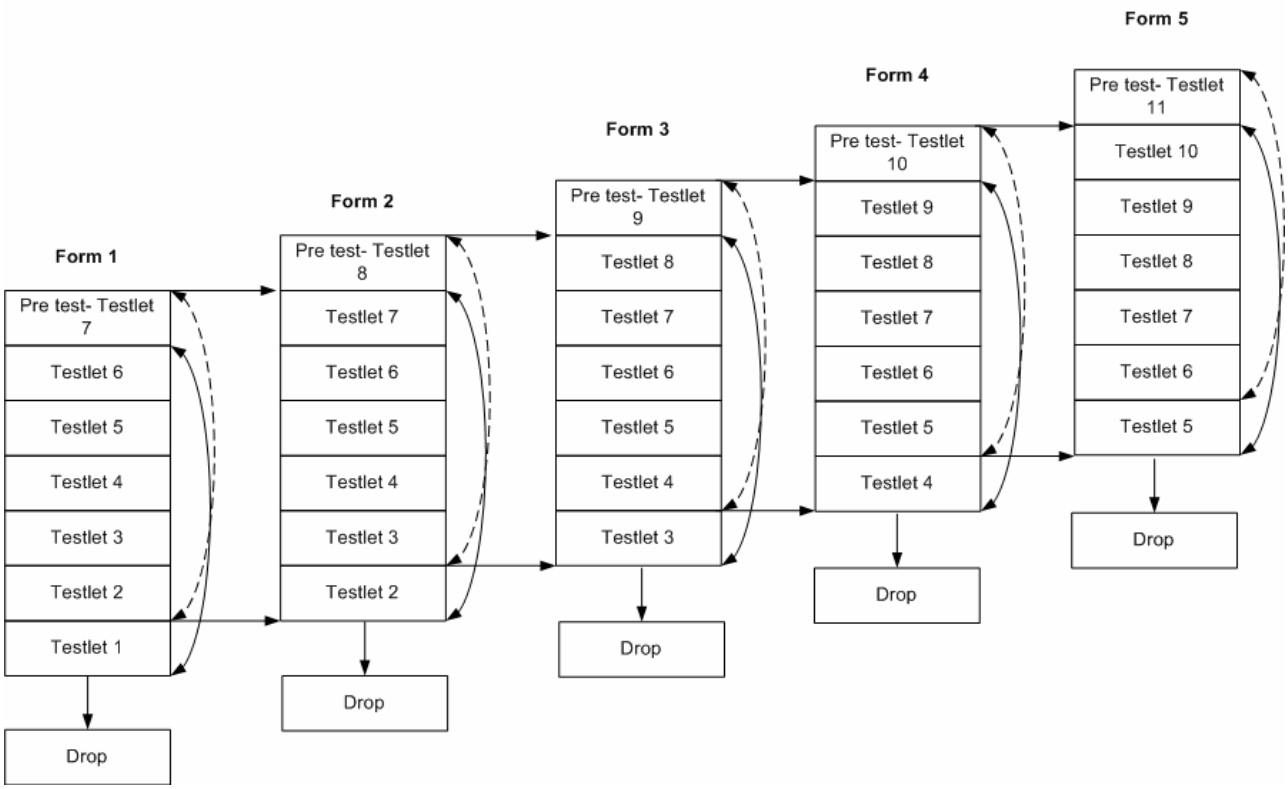

Figure 2. **A hypothetical testing design for the current study.**

As evident, there can be two equating conversions for Form 5, one of which is derived via the chain of equatings described above. One can also conduct a separate equating of Form 5 using the NEAT design, where Form 5 is equated directly to Form 1 using common items (Note that Testlets 5 and 6 are common to Forms 1 and 5 and therefore can be used as an anchor to equate Form 5 directly to Form 1). These two equatings can be compared with a criterion to evaluate which design performs better under small sample conditions.

Although the SiGNET design involves several intermediate SG equatings, which have the lowest random equating error compared to other equating methods, these errors can accumulate over time and may eventually become large enough to cause a concern. Therefore, this study also assessed whether using longer equating chains under the SiGNET design substantially increased random equating error and bias compared to the NEAT design, where the new form is equated directly to an old form and avoids the long chain of equatings.

Note that one could use Form 1 as the old form whenever a new form under the SiGNET design is created. For example, Form 2 can be equated to Form 1 and Form 3 can also be equated to Form 1. This may help avoid accumulating error and bias that can be caused by the long chain of equatings (i.e., Form 5 to Form 1 via Forms 4, 3, and 2). However, the long chain of equatings has a potential advantage, in that it allows testing programs to eventually reach a stage where a new form (e.g., Form 8) has no overlap with the initial form (i.e., Form 1). Once the testing program has created several such forms, they can be alternated in different administrations to lower the risk of item exposure. Moreover, tests evolve over time because content areas that are deemed more important during one time period may become less important with time. Thus using Form 1 as the base form for all future equatings may not be reasonable.

One can also make similar comparisons for Forms 4, 3, and 2, where they are linked to Form 1 via equating chains or directly equated to Form 1 using common items. As is evident in Figure 2, the NEAT equating of Form 2 to Form 1 will have the largest number of testlet anchors (overlap) as compared to other NEAT equatings (e.g., Form 3 to Form 1 or Form 4 to Form 1). The current study also compared equatings of Form 2 to Form 1 using either a NEAT design or a SiGNET design. This comparison should be informative, because if equating Form 2 to Form 1 with a very large anchor set resulted in a substantial improvement in equating for some small sample conditions (e.g., $N \geq 50$), then moving to the SiGNET model in those situations may not be needed. Hence, this study compared the SiGNET and NEAT equatings under long and short equating chain conditions. Throughout this paper, the long equating chain condition will refer to the Form 5 to Form 1 equating via Forms 4, 3, and 2, and the short chain condition will refer to the Form 2 to Form 1 equating with no intermediate equatings in between.

Deriving a criterion equating function. When evaluating a particular equating design against another equating design, it is desirable to compare both designs with a criterion equating (i.e., the equating relationship in the population). Although the true criterion is impossible to know, a proxy of the true criterion can be obtained under certain conditions. This study was designed to create a hypothetical situation where a reasonably good criterion could be obtained and results of the SiGNET and NEAT designs for small sample conditions could be evaluated against this criterion. As seen in Figure 2, for the Form 5 to Form 1 equating comparisons, since both Form 5 and 1 are taken by the total group of 23,000 examinees, an equating conversion for Form 5 can be obtained by equating Form 5 directly to Form 1 using an SG equating design. By

using the SG design with such a large data set, one can be fairly confident that the resulting conversion is a very good approximation of the equating in the population (the true criterion). Similarly, for the Form 2 to Form 1 equating comparison, since both Form 2 and 1 are taken by the total group of examinees, an equating conversion for Form 2 can be obtained by equating Form 2 directly to Form 1 using an SG equating design. This equating served as the criterion for the Form 2 to Form 1 comparisons.

Sample Size Conditions

For the small sample NEAT equatings, four small sample size conditions were chosen in this study (i.e., N = 10, 15, 25, and 50). Although other small sample size conditions can be studied, it was decided to use these sample sizes because they seemed small enough to create problems in equating (especially sample sizes of 10, 15, and 25) and would likely benefit the most under the SiGNET design. Previous research has also shown that equating is problematic with these sample sizes (see Kim et al., 2006; Skaggs, 2005).

It is crucial to create a balance between conducting a good equating with sufficient data and not overexposing a particular test form by repeated administrations in an effort to collect sufficient data. Therefore, SiGNET equatings of sample sizes 50, 100, and 150 were used to provide some guidelines as to when one can consider having enough data to equate so as to prevent overexposing a particular test form. Consider the N = 15 sample size condition as an example. It would take approximately 3, 7, and 10 administrations to accumulate the SG sample sizes mentioned above (i.e., N = 50, 100, and 150). Therefore, if the N = 50 sample provided a similar equating as the N = 100 or 150 sample sizes, then it may be preferable to equate as soon as the accumulated data reach a sample size of 50 to prevent overexposure of the particular form.

Although there is no direct relationship between the NEAT sample sizes and the SiGNET sample sizes, it is reasonable to expect the SiGNET sample sizes in a real testing situation to be larger than the NEAT sample sizes, because the SiGNET design allows for the accumulation of samples over successive test administrations. This is not possible with the NEAT design. Under the NEAT design, if the sample size is very small during the first administration of a new form, then equating has to be conducted with the very small sample unless one chooses not to equate at all, which may have other adverse implications. After the sample size conditions were chosen, the study was conducted for the long chain condition (Form 5 equated to Form 1) using the steps

described below (refer to Figure 2 as necessary). The same steps were followed for the short chain condition (Form 2 equated to Form 1).

Step 1. To estimate the criterion equating function, the mean and standard deviation (*SD*) of the group of examinees on Form 5 were set equal to the mean and *SD* of the group of examinees on Form 1 (i.e., a direct linear equating) using an SG equating design. The resulting conversion was considered as the criterion to which the small sample SiGNET and NEAT[2] equatings would be compared. Note that a nonlinear (i.e., equipercentile) equating could also be derived under this SG equating design and used as the criterion. But since the equipercentile conversion looked reasonably linear, especially in the middle of the distribution where most of the data were clustered, it was decided to keep the linear conversion as the criterion. Furthermore, since the small sample equating was to be conducted using linear equating (equipercentile equating with very small samples often results in a very large error), it was deemed appropriate to use the linear equating as the criterion.

Step 2. For a particular sample size condition (e.g., $N = 10$), sample sizes of 10 each were drawn without replacement for Form 5 and Form 1 from the full data set, and a NEAT equating was conducted using these samples to equate Form 5 directly to Form 1. Then for the SiGNET equatings under this sample size condition, sample sizes of 50 each were drawn without replacement from the full data set for Forms 1-5. A chain of equatings was completed, whereby a conversion for Form 2 was created using an SG equating with Form 1 data, a conversion for Form 3 was created using an SG equating with Form 2 data, conversion for Form 4 was created using an SG equating with Form 3 data, and finally a conversion for Form 5 was created using an SG equating with Form 4 data. This chain of equating produced a conversion for Form 5. This procedure was repeated 200 times. In the sample selection procedure described above, examinees were randomly selected without replacement within a replication (i.e., one small sample run) but with replacement among the 200 replications. The procedure described above was then repeated for the remaining sample size conditions (i.e., $N = 15$, 25, and 50 along with their respective SiGNET equatings of $N = 50$, 100, and 150 within each sample size condition). The procedure described above was also repeated for the short chain condition (i.e., Form 2 equated directly to Form 1). Since there are no intermediate equatings in between the Form 2 to Form 1 equating, the only difference from the step described above is that small sample draws from the full data for any intermediate forms were not required. Therefore, in total there were 12 conditions for the

long equating chain condition and 12 conditions for the short equating chain condition (see Table 1, which presents all of the conditions used in this study).

Table 1

Conditions Used in the Study for the Long and Short Equating Chains

Sample size conditions	Comparisons[a]		
N = 10	NEAT (*N* =10) vs. SiGNET (*N* = 50)	NEAT (*N* =10) vs. SiGNET (*N* = 100)	NEAT (*N* =10) vs. SiGNET (*N* = 150)
N = 15	NEAT (*N* =15) vs. SiGNET (*N* = 50)	NEAT (*N* =15) vs. SiGNET (*N* = 100)	NEAT (*N* =15) vs. SiGNET (*N* = 150)
N = 25	NEAT (*N* =25) vs. SiGNET (*N* = 50)	NEAT (*N* =25) vs. SiGNET (*N* = 100)	NEAT (*N* =25) vs. SiGNET (*N* = 150)
N = 50	NEAT (*N* =50) vs. SiGNET (*N* = 50)	NEAT (*N* =50) vs. SiGNET (*N* = 100)	NEAT (*N* =50) vs. SiGNET (*N* = 150)

Note. NEAT = nonequivalent groups with anchor test, SiGNET = single group nearly identical tests.

[a] 200 equating runs conducted for each comparison.

As evident, the sample sizes for the SiGNET equatings are mostly larger than the sample sizes for the NEAT equatings, and therefore the comparison of SiGNET with NEAT equatings may seem unfair (i.e., since SiGNET has a larger sample size, one would expect a smaller equating error compared to the NEAT equatings). However, as mentioned earlier, because the SiGNET design allows for the accumulation of data before equating, it seemed practical to compare SiGNET equatings based on larger samples to NEAT equatings based on smaller samples.

Step 3. For the small sample conditions, both variability and accuracy of the equatings under the SiGNET and NEAT designs were evaluated using different statistical indices. To provide a measure of variability, the standard deviation of the 200 replications (i.e., equated

scores under the NEAT and SiGNET designs) was calculated for each score point. This is the conditional standard error of equating (CSEE), and the formula is

$$CSEE_j = \sqrt{\frac{1}{I}\sum_i \left(\hat{e}_y(x_{ij}) - \frac{\sum_i \hat{e}_y(x_{ij})}{I} \right)^2} \ ,$$

where I is the number of replications in the simulation (i = 1 to I, and I = 200) and $\hat{e}_y(x_{ij})$ is the NEAT or single group *X-to-Y* equated score at score = x_j estimated for replication = i of a particular sample size. An average of the CSEEs was also calculated to get an estimate of overall standard error of equating (SEE). The formula is

$$AvgSEE = \sqrt{\sum_j p_j CSEE_j^2} \ ,$$

where p_j is the raw proportion of examinees at score = x_j in the total population data. Multiplying the CSEE by p_j ensures that CSEEs are appropriately weighted in the calculation of the average SEE. To provide a measure of accuracy, a bias statistic for each score point (i.e., conditional bias) was calculated. The formula is

$$CBias_j = \frac{1}{I}\sum_i \left(\hat{e}_y(x_{ij}) - e_y(x_j) \right),$$

where I is the number of replications in the simulation (i = 1 to I, and I = 200), $e_y(x_j)$ is the criterion equipercentile or linear *X-to-Y* single group equated score at score = x_j estimated with the total population data, and $\hat{e}_y(x_{ij})$ is the NEAT or single group *X-to-Y* equated score at score = x_j estimated for replication = i of a particular sample size. An average of the bias values for all score points was also calculated to get an estimate of the overall equating bias across all score points. The formula is

$$Bias = \sum_j p_j CBias_j \ ,$$

where p_j is the raw proportion of examinees at score = x_j in the total population data. Although the bias statistic cancels out positive and negative differences around the criterion, it is helpful in summarizing whether, on an average, estimates derived using SiGNET or NEAT design deviate from the criterion in either the positive or negative direction. An overall bias statistic that is close to zero would indicate a small bias. Finally, the average root mean squared deviation (RMSD) was also calculated for the NEAT and SiGNET equatings. The formula is

$$AvgRMSD = \sqrt{AvgBias^2 + SEE^2} \ ,$$

where $AvgBias^2$ is the sum of the squared conditional bias values weighted by the raw proportion of examinees at each score point. The formula is

$$AvgBias^2 = \sum_j p_j CBias_j^2 \ .$$

The average RMSD is a useful statistic because it provides an estimate based on combining information from systematic error (bias) and random error (SEE).

The Difference That Matters (DTM) Criterion

When comparing the small sample NEAT and the SiGNET equating, a smaller value for the statistics stated above would indicate a better equating in terms of lower standard error, or bias, or both. One practical guideline to evaluate different equatings is the use of the difference that matters (DTM; Dorans & Feigenbaum, 1994). Briefly stated, Dorans and Feigenbaum (1994) defined a DTM as any score difference that would make a difference in score reporting once scores were rounded. In this study, where scores progressed in 1-point increments, the DTM was defined as any score difference that is equal to or greater than 0.5.

Using a DTM of 0.5 as a criterion to assess the average SEE, bias, and average RMSD seemed reasonable because, if on average the error, bias, or RMSD for several small sample equating replications is less than 0.5, then it essentially means that the difference between the small sample equatings and the criterion is not large enough to have any practical impact on the reported scores. However, it should be noted that these are average statistics, and there may still be scores from certain replications that are above 0.5 and may make a practical difference.

Results

Overall accuracy and variability of equatings under small sample NEAT and SiGNET designs were estimated using the average SEE, bias, and average RMSD indices. These results are presented in Tables 2 and 3. In addition to the average accuracy and variability results, conditional standard error and bias estimates were calculated and are presented in Figures 3–6. Results from the Form 5 to Form 1 equatings are presented first, followed by results from the Form 2 to Form 1 equating.

Table 2

Average Squared Bias and Error and Mean-Squared Deviation for the Nonequivalent Groups With Anchor Test (NEAT) Versus Single Group Nearly Identical Tests (SiGNET) Designs (Form 5 to Form 1)

Accuracy and variability indices	NEAT equatings				SiGNET equatings		
	N = 10	*N* = 15	*N* = 25	*N* = 50	*N* = 50	*N* = 100	*N* = 150
Average SEE	2.59	2.05	1.62	1.27	0. 83	0.56	0.47
Bias	-0.19	0.10	-0.02	-0.06	-0.03	0.02	0.01
Average bias^2	0.05	0.02	0.03	0.00	0.00	0.00	0.00
Average RMSD	2.60	2.05	1.63	1.27	0.83	0.56	0.47

Note. RMSD = root mean squared deviation, SEE = standard error of equating.

Table 3

Average Squared Bias, Error, and Mean Squared Deviation for the Nonequivalent Groups With Anchor Test (NEAT) Versus Single Group Nearly Identical Tests (SiGNET) Designs (Form 2 to Form 1)

Accuracy and variability indices	NEAT equatings				SiGNET equatings		
	N = 10	*N* = 15	*N* = 25	*N* = 50	*N* = 50	*N* = 100	*N* = 150
Average SEE	0.85	0.65	0.53	0.42	0.42	0.27	0.22
Bias	-0.01	0.02	-0.01	-0.01	0.01	0.03	-0.01
Average bias^2	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Average RMSD	0.86	0.65	0.53	0.42	0.42	0.27	0.22

Note. RMSD = root mean squared deviation, SEE = standard error of equating.

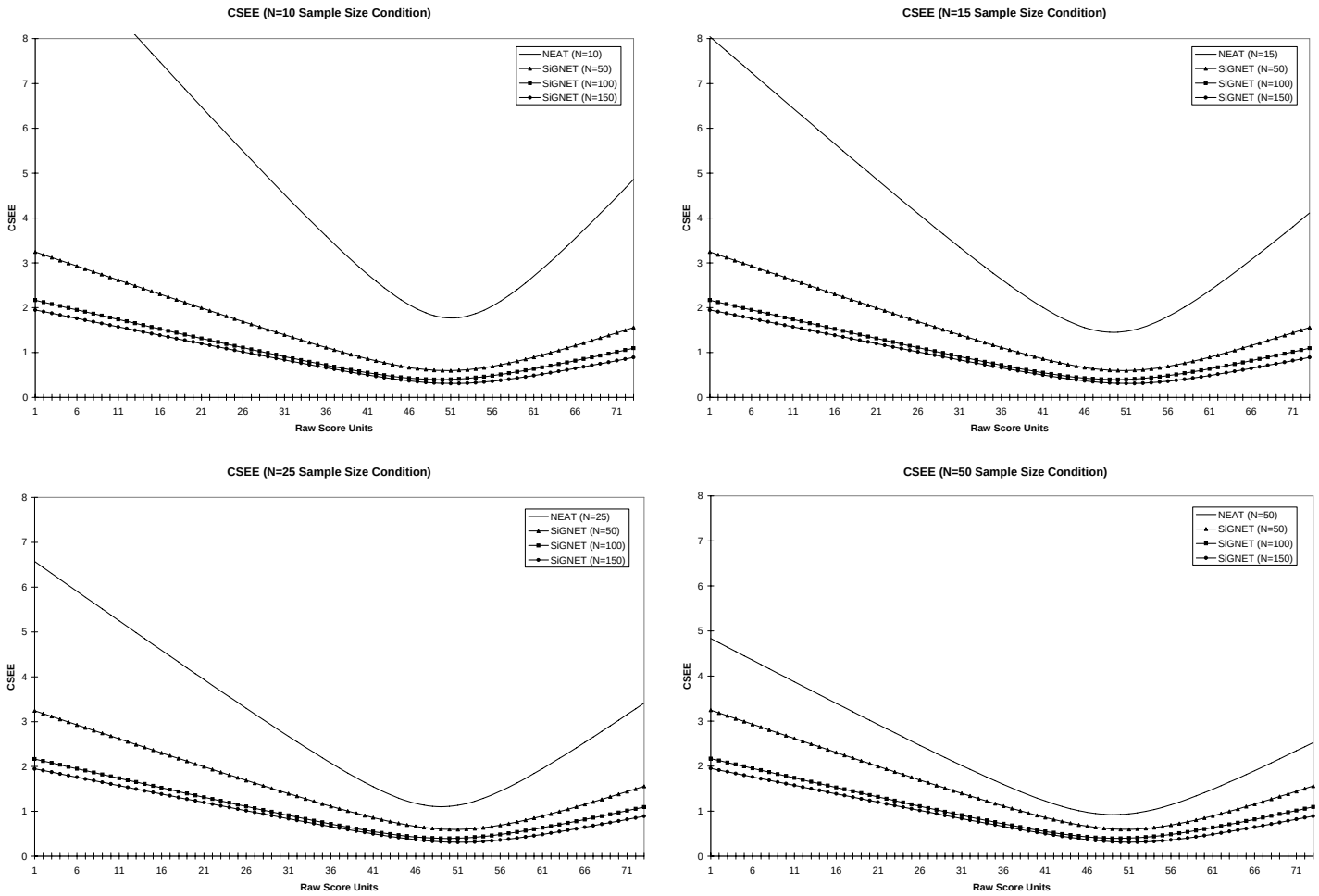

Figure 3. **Conditional standard error of equating for the Form 5 to Form 1 equatings for sample sizes of 10, 15, 25, and 50.**

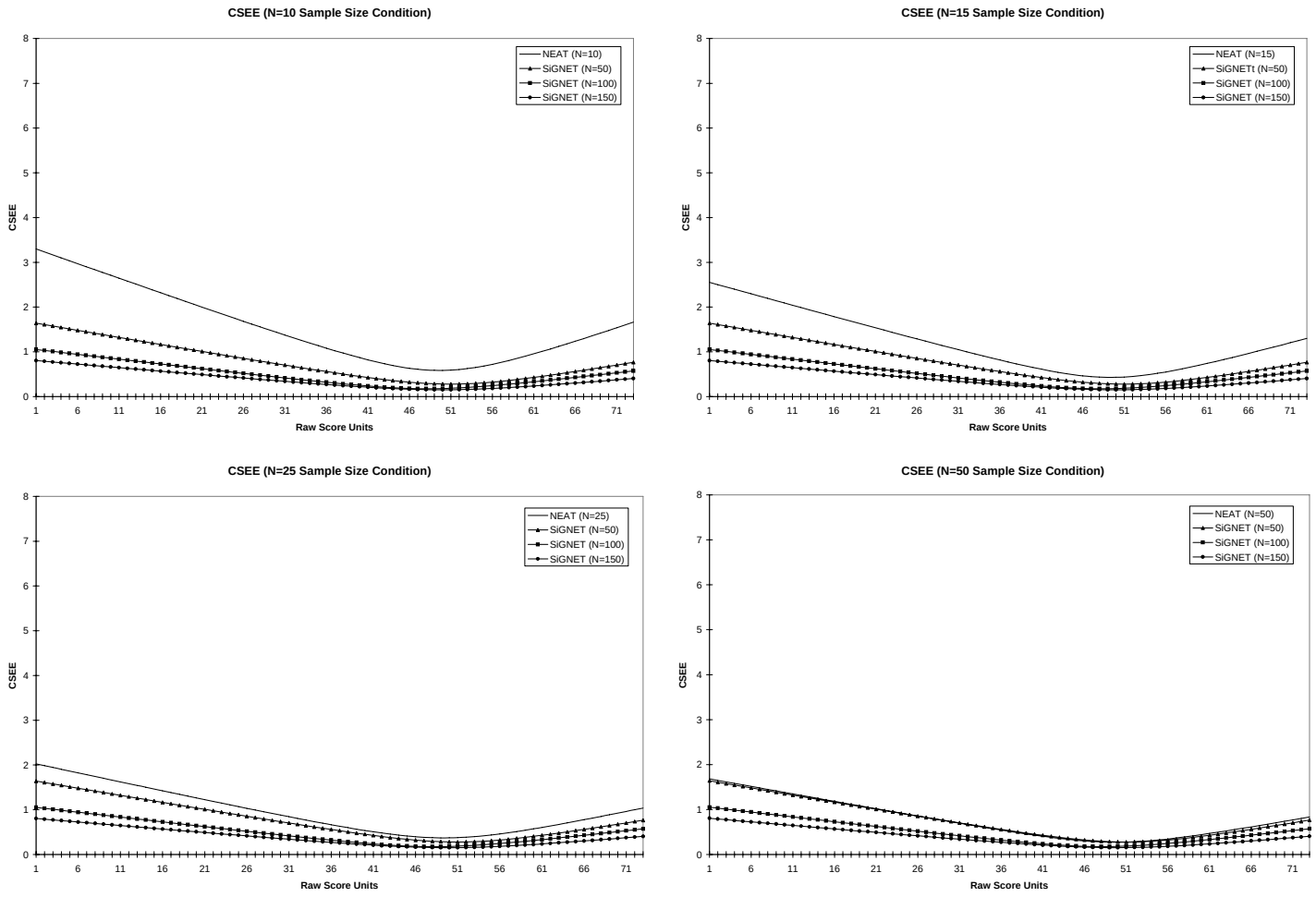

Figure 4. **Conditional standard error of equating for the Form 2 to Form 1 equatings for sample sizes of 10, 15, 25, and 50.**

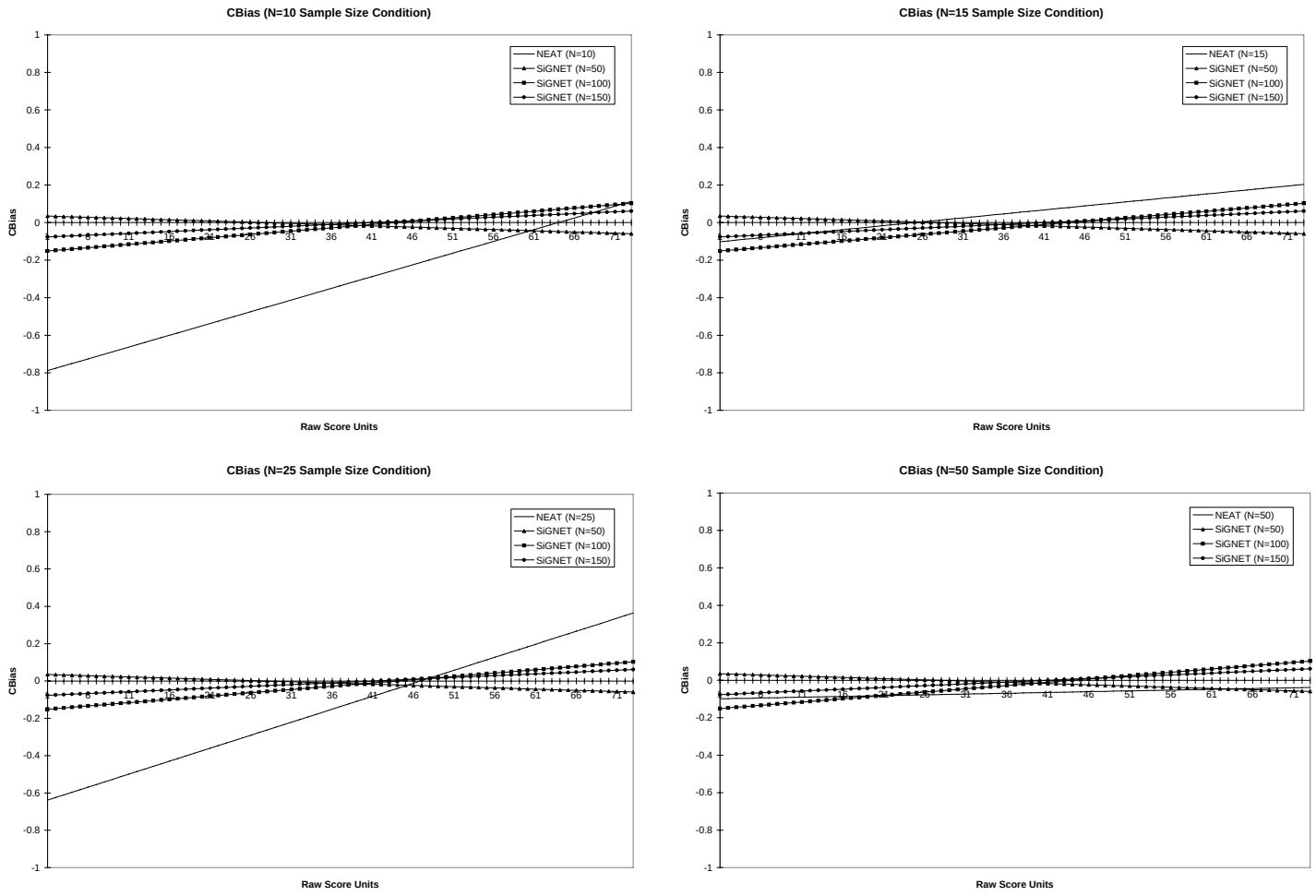

Figure 5. **Conditional bias for the Form 5 to Form 1 equatings for sample sizes of 10, 15, 25, and 50.**

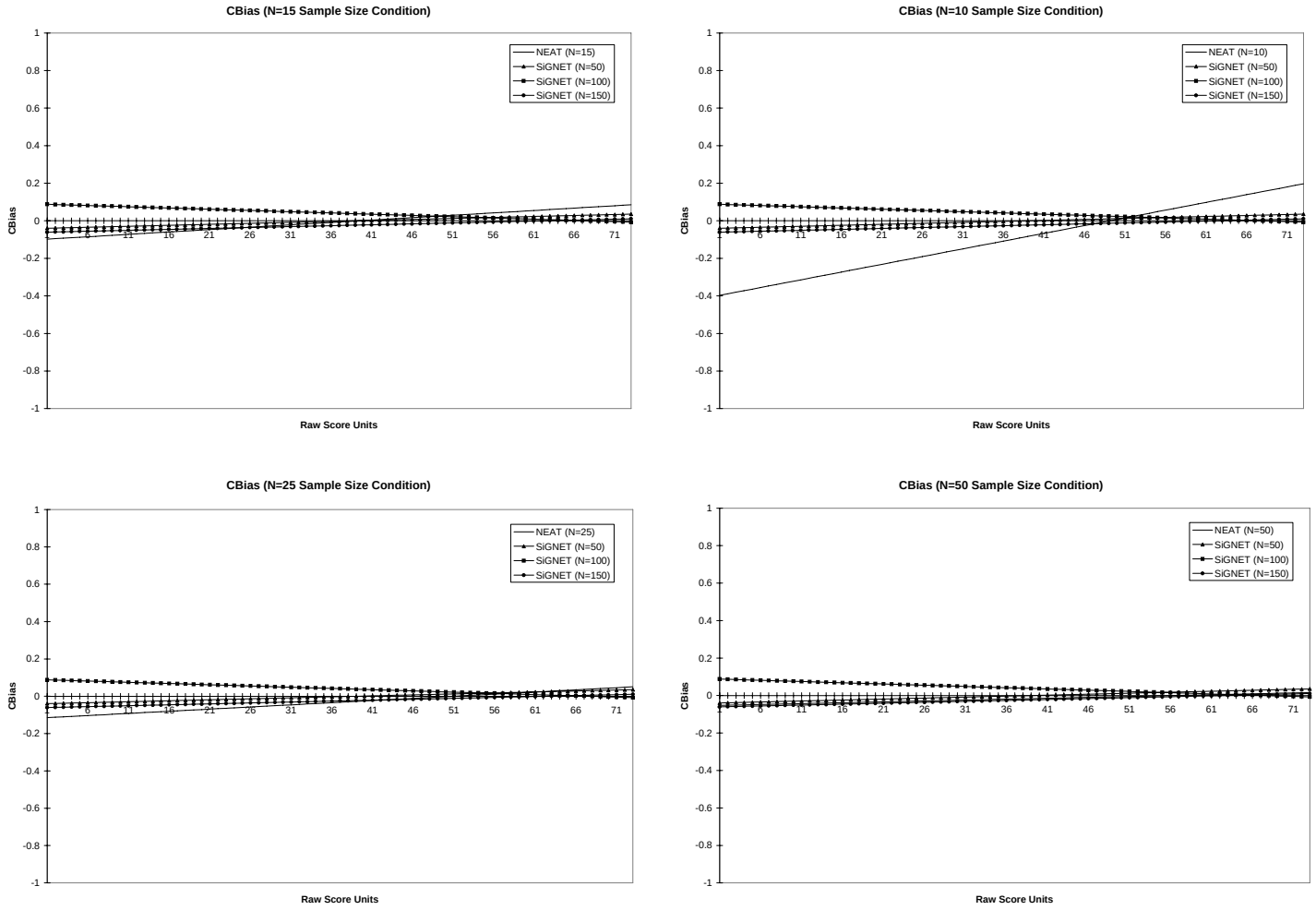

Figure 6. **Conditional bias for the Form 2 to Form 1 equatings for sample sizes of 10, 15, 25, and 50.**

Results From the Form 5 to Form 1 Equatings

Average standard error of equating (SEE), bias, and average root mean squared deviation (RMSD) results. For the sample size condition with 10 test takers, the average SEE for the NEAT equating was 2.59 (see Table 2). This is much larger than the average SEE for SiGNET equatings based on sample sizes of 50, 100, and 150, where the average errors were 0.83, 0.56, and 0.47, respectively. The bias for the NEAT equating under this sample size condition was -0.19. This is also larger than the bias for the SiGNET equatings based on sample sizes of 50, 100, and 150, where the bias values were -0.03, 0.02, and 0.01, respectively. Finally, the average RMSD for the NEAT equating was 2.60, which is much larger than the average RMSDs for the SiGNET equatings based on sample sizes of 50, 100, and 150, where the average RMSD values were 0.83, 0.56, and 0.47, respectively.

As seen in Table 2, the average SEEs for the remaining sample size conditions followed a similar trend to the finding for the sample size condition with 10 test takers (i.e., the NEAT average SEEs were larger than the SiGNET average SEEs). The bias estimates for both the NEAT and SiGNET equatings were rather small, although the bias values for all the SiGNET equatings were slightly smaller than the bias values estimated for the N = 10 and N = 15 NEAT equatings. Finally, the average RMSD values for the remaining NEAT equatings (N = 15, 25, and 50) were larger than the SiGNET average RMSDs for sample sizes of 50, 100, and 150.

The average SEEs and RMSDs for the NEAT equatings were largest for the sample of 10 test takers and became progressively smaller for the larger sample sizes (see Table 2). This was also true for the SiGNET equatings. For the SiGNET equatings, although the average error and RMSD decreased as sample size increased, they decreased more when sample size increased from 50 to 100, compared to the sample size increase from 100 to 150.

Note that although there are separate average SEE, bias, and average RMSD values for the NEAT equatings for each sample size condition, only a single set of values for the SiGNET equatings is presented in Table 2 (N = 50, 100, and 150). Even though the values for the SiGNET equatings (N = 50, 100, and 50) were not expected to be identical under the different sample size conditions because they were based on 200 different replications in each sample size condition, the actual values in the current study were very similar. Therefore, for economy of presentation, we reported the values for the SiGNET equatings that were estimated in the N = 10 NEAT condition. This is also true for the results in the next section (i.e., Form 2 to Form 1 equatings).

Conditional Standard Errors and Bias

For tests that employ cut scores, the CSEEs and conditional bias values may be more informative than the average error or bias. For this study it was decided to examine CSEEs and conditional bias values for all score points within ±1.5 *SD* from the mean. The ±1.5 *SD* from the mean approach has also been used in past research to evaluate CSEEs (see Parshall, Du Bose, Houghton, & Kromrey, 1995). Moreover, for licensure tests, cut scores typically fall within this range, and therefore it seemed reasonable to use this range when examining the CSEEs and conditional bias values. For this study it was decided to use the mean and *SD* for the new form in the total data to calculate this range. Considering the mean and *SD* for the new form in the total data (i.e., Form 5) which were 47.73 and 8.83, respectively, the minimum and maximum scores for ±1.5 *SD* from the mean are roughly 35 and 61. For the Form 2 to Form 1 equatings, this range was also calculated, and the minimum and maximum scores in this range are 33 and 61. These ranges will be considered when evaluating the CSEEs and conditional bias values throughout this paper.

As seen in Figure 3, for the sample size condition with 10 test takers, the CSEEs for the NEAT equatings were larger than the CSEEs for the SiGNET equatings based on sample sizes of 50, 100, and 150. Similarly, the conditional bias estimates for the NEAT equatings were mostly larger than the SiGNET equatings (see Figure 5). This trend was evident in the remaining sample size conditions (i.e., N = 15, 25, and 50) where the SiGNET equatings had lower CSEEs and conditional bias than the NEAT equatings.

Similar to the average statistics described earlier, the CSEEs for the NEAT equatings were largest for the sample size condition with 10 test takers and became progressively smaller for the larger sample sizes. This was also true for the conditional bias, except for one condition (i.e., N = 25) where the conditional bias values were slightly larger than those for the sample size condition with 15 test takers for several score points. However, the CSEEs at those points were also large, suggesting that the observed difference in bias in these two conditions may be attributed to random variability. For example, the conditional bias values for the sample size conditions with 15 or 25 test takers were -0.26 and 0.01, respectively at Raw Score 29. However the difference between two conditional bias values is 0.27, which is much smaller than the CSEE at Score Point 29 from either the sample size condition with 15 or 25 test takers, where the CSEEs were 3.79 and 3.04, respectively, indicating that the difference may be attributed to

random error. The CSEEs were also the smallest around the mean of the score distribution ($\bar{X}_{Form1}$ = 46.80 and $\bar{X}_{Form5}$ = 47.73 in the total sample), and the error increased monotonically (but not linearly) as one progressed toward the tails of the distributions. Finally, considering the DTM criteria and the *±1.5 SD* from the mean range, one can see that the CSEEs for the NEAT equatings for all sample size conditions were larger than a DTM of 0.5 for the 35 to 61 score range and beyond, indicating that for tests employing cut scores using NEAT equating can potentially result in equating errors that are not small enough to ignore[3] at those score points.

The CSEE for the SiGNET equatings decreased as sample size increased. As seen in Figure 3, the CSEE for the sample size condition with 50 test takers was larger than the CSEEs for the other two sample size conditions. The difference in the CSEEs for the samples of 100 and 150 test takers is small, especially around the mean of the distribution, where there are more data than in the tails of the distribution. The conditional bias values for the SiGNET equatings were quite small for the different sample size conditions. However, for the samples of 100 and 150 test takers, the bias values were more similar compared to the sample size condition with 50 test takers, where the bias values showed an opposite trend (see Figure 5). Similar to the NEAT equatings, the CSEEs were also the smallest around the mean of the score distribution, and the error increased monotonically (but not linearly) as one progressed toward the tails of the distributions. However the increase in error towards the tails was not as large as in the NEAT equatings because of larger data accumulated in case of the SiGNET equatings. As seen in Figure 3, although the CSEEs for the SG equatings were smaller than the NEAT equatings, if one considers the score range of 35 to 61, some of the CSEEs within this range were still larger than 0.5, even for the SG equatings with 150 test takers. The CSEEs were smaller than 0.5 for all score points within the 40 to 60 score range, but not beyond that range.

At this stage, it seemed reasonable to run additional SiGNET equatings with larger sample sizes to evaluate when the CSEEs were smaller than 0.5 for all scores within the 35 to 61 score range. Results from these analyses indicated that with sample size of 250, the SiGNET equatings produced CSEEs lower than 0.5 in this score range (see Figure A1 in the appendix). The conditional bias estimates were less than 0.5 for all score points. An additional run was also conducted on sample size of 200, and those results indicated that the CSEEs were lower for most of the score points within the 35 to 61 score range, but they were slightly higher than 0.5 for some scores close to 35.

Results From the Form 2 to Form 1 Equatings (i.e., the Short Chain)

Average standard error of equating (SEE), bias, and average root mean squared deviation (RMSD) results. As seen in Table 3, for the sample size condition with 10 test takers, the average SEE for the NEAT equating was 0.85. This is larger than the average error for SiGNET equatings based on sample sizes of 50, 100, and 150, where the average errors were 0.42, 0.27, and 0.22, respectively. The bias for the NEAT equating under this sample size condition was very close to zero. The bias values for the SiGNET equatings based on sample sizes of 50, 100, and 150 were also close to zero. Finally, the average RMSD for the NEAT equating was 0.857, which is larger than the average RMSDs for the SiGNET equatings based on sample sizes of 50, 100, and 150, where the average RMSD values were 0.42, 0.27, and 0.22, respectively.

As seen in Table 3, the average SEEs for the remaining sample size conditions followed a trend that is similar to that for the sample size condition with 10 test takers. However, the difference between the NEAT and SiGNET equatings became small as the sample got larger. For example at sample size 50, the average SEEs for the NEAT and SiGNET equatings were almost identical. This was also true for the average RMSD values. Finally, the bias values for both the NEAT and SiGNET equatings were close to zero for the remaining sample size conditions. Similar to the Form 5 to Form 1 equating, the average SEEs and RMSDs for the NEAT equatings were largest for the sample size condition with 10 test takers and became progressively smaller for the larger sample sizes. This was also true for the SiGNET equatings. As observed in the previous condition, the average SEEs and RMSDs decreased more when the sample size increased from 50 to 100 compared to when the sample size increased from 100 to 150.

Conditional standard errors and bias. As seen in Figure 4, for the sample size condition with 10 test takers, the CSEEs for the NEAT equatings were larger than the CSEEs for the SiGNET equatings based on sample sizes of 50, 100, and 150. The trend was similar for the sample size condition with 15 test takers and to some extent for the sample size condition with 25 test takers. But for the sample size condition with 50 test takers, there was virtually no difference in the CSEEs for the NEAT or SiGNET equatings. For the sample size condition with 10 test takers, the conditional bias for most score points was larger for the NEAT than for the SiGNET equatings. The conditional bias values for the remaining sample size conditions were about the same (see Figure 6). Similar to the Form 5 to Form 1 equatings, the CSEEs were the

smallest around the mean of the score distribution ($\bar{X}_{Form1}$ = 46.80 and $\bar{X}_{Form2}$ = 47.12 in the total sample), and the error increased monotonically (but not linearly) as one progressed toward the tails of the distributions.

The CSEEs for the NEAT equatings for all sample size conditions except the N = 50 sample size condition were larger than a DTM of 0.5 for several score points in the 33 to 61 score range and beyond. However for the sample size condition with 50 test takers, the CSEEs were lower than the DTM for all scores points except scores 33, 34, and 35, where the CSEEs were slightly larger than 0.5. This was also true for the SiGNET equating with 50 test takers. For the SiGNET equatings with 100 and 150 test takers, the CSEEs were smaller than the DTM for the 33 to 61 score region.

Discussion and Conclusion

Small samples are more likely to differ from the population and therefore are more likely to produce equating results that are less stable. As mentioned earlier, for many testing programs, small samples are a reality. However, when new forms are introduced, they must be equated to ensure comparability with scores on other existing forms of that test. Therefore, the purpose of this study was to compare a small sample equating design known as the SiGNET equating design with the NEAT design in small sample conditions, to evaluate whether using the SiGNET equating design leads to an improvement in equating over that for the NEAT equating design. The accuracy of equating in both the SiGNET and NEAT designs was evaluated by comparing these equatings with a criterion equating that was derived from the total available data using an SG equating design. The standard deviation of the 200 replications of the small sample equatings provided an indication of random sampling error. The average of the difference between the small sample SiGNET or NEAT equatings from the criterion provided an indication of bias. Overall, the SiGNET equating design resulted in a better equating (i.e., less sampling error and bias) compared to the NEAT design under different small sample size conditions.

Findings from the Form 5 to From 1 equatings are summarized first: (a) for all sample size conditions, the average SEE under the SiGNET equating design was smaller compared to that for the NEAT design; (b) for all sample size conditions, the average bias was small (close to zero) for the small sample equatings under both the SiGNET and NEAT equating designs; (c) for all sample size conditions, the average RMSD was smaller for the SiGNET compared to the

NEAT equating design; (d) for all sample size conditions, the CSEE was smaller for the SiGNET (including the additional runs with sample sizes of 200 and 250) compared to the NEAT equating design; and (e) for all sample size conditions, the conditional bias values were small (less than 0.5) for the SiGNET equatings. The NEAT equatings were small for most points on the score scale, except for sample sizes of 10 and 25, where they were slightly higher for the lower scores in the score distribution.

These results, when considered in light of the DTM, suggest that NEAT equatings are problematic for all small sample size conditions (i.e., N = 10, 15, 25, and 50). Using the SiGNET equating design results in an improvement over the NEAT equatings. However, we found that the average SEE was still larger than the DTM of 0.5 for sample sizes of 50, and quite close to 0.5 for sample sizes of 100. Using the DTM criteria, all the bias estimates for the SiGNET and NEAT equatings can be considered to be very small. Finally, the average RMSD was larger than the DTM for all the NEAT equatings, and it was larger than the DTM for the SiGNET equatings for the sample size condition with 50 test takers, but very close to the DTM for the sample size conditions with 100 and 150 test takers.

The CSEEs were also larger for the NEAT equatings compared to the SG equatings. They were also larger than the DTM for the NEAT equatings in all sample size conditions. For the SiGNET equatings (i.e., N = 50), the CSEEs were larger than the DTM for most score points, except for the middle score points, where they were very close to the DTM. This was also true for sample sizes of 100 and 150. However for these two sample size conditions, the CSEEs were lower than the DTM for a wider range of the score points around the middle of the distribution. Because the CSEEs for the 150 sample size condition were not lower than 0.5 for all score points in the 35 to 61 score range (i.e., ±1.5 *SD* from the mean), additional SiGNET equatings with sample sizes of 200 and 250 were conducted and revealed that, for the sample size condition with 250 test takers, the CSEEs were less than 0.5 for all score points between 35 and 61.

Based on these findings, a preliminary recommendation is to accumulate at least 250 examinees before conducting the SiGNET equating. This seems reasonable, because for this sample size the average SEE and RMSD (which were both 0.336) for the SiGNET equating were smaller than the DTM, and the CSEEs and conditional bias were also smaller than the DTM[4] for all score points at ±1.5 *SD* from the mean. The additional analyses for the NEAT equatings[3]

suggest that at least 600 examinees are required to get a trustworthy NEAT equating (with approximately 25% content overlap between new and old forms) in terms of lower CSEEs.

Form 2 to From 1 equatings are summarized as follows: (a) for all sample size conditions, the average CSEE under the SiGNET equating design was smaller compared to that for the NEAT design; (b) for all sample size conditions, the average bias was small (close to zero) for the small sample equatings under both the SiGNET and NEAT designs; (c) for all sample size conditions, the average RMSD under the SiGNET equating design was smaller compared to the NEAT design; (d) for all except the sample size condition with 50 test takers, the CSEE under the SiGNET equating design was smaller compared to that for the NEAT design. The CSEEs for the SiGNET and NEAT equatings for sample sizes of 50 were about the same; and (e) the conditional bias values were small for all sample size conditions studied. Although the conditional bias for the sample size condition with 10 test takers was slightly larger around the tails of the distribution, it was smaller than the DTM. These results were also considered in light of the DTM and suggested that NEAT equatings were problematic (in terms of random error) for all except the sample size condition with 50 test takers, where the average SEE was smaller than the DTM. The overall bias for all conditions was close to zero, and the average RMSD showed a similar trend to the average SEE. The average SEE, bias, and average RMSD for all the SiGNET equatings were small compared to the DTM. The CSEEs were larger than the DTM for the sample size condition with 10 test takers for the NEAT equatings. For the other NEAT equatings with the larger sample sizes, the CSEEs were closer to or smaller than the DTM around the middle of the distribution. However, for the NEAT equatings with 50 test takers, the CSEEs and conditional bias were less than the DTM for most scores between 33 and 61 (±1.5 *SD* from the mean). Similarly, for the SiGNET equatings, the CSEEs and conditional bias were smaller than the DTM for most scores between 33 and 61.

Considering these results, a preliminary recommendation is to use the SiGNET model for sample sizes smaller than 25 in situations where the amount of content overlap (i.e., anchor item set) between the new and old forms is substantially large (over 80 %). For sample sizes of 50 or more examinees, both the NEAT and SiGNET designs perform equally well and either may be used, although the NEAT design may be a slightly better option for limiting item exposure.

In practice, whether one can meet the sample size requirements recommended above would depend on a number of factors, such as the availability of such sample sizes (and whether

it is possible to obtain them), risk of exposure versus the need to accumulate sufficient number of examinees to conduct a SiGNET equating, probable consequences of reporting scores that have a modest amount of error (e.g., CSEEs slightly over 0.5 but not higher than 1), and the actual impact of larger CSEEs or bias on pass and fail status of students (particularly for licensure tests). For licensure tests where most scores fall in the middle or right side of the score distribution, it may not matter much if the CSEEs at a lower score region are slightly higher than the DTM, because in those cases, most examinees would likely pass if a cut score was set reasonably low.

In summary, these results suggest that the SiGNET equating design may be preferred under small sample conditions, because using this design makes it possible to conduct an SG pre-equating of a new form with the added benefit of accumulating data when needed, thereby leading to an improved equating compared to NEAT equatings. However, if the amount of overlap is substantially large between the old and new forms and the sample size is about 50 or higher, then the regular NEAT design is a reasonable option. In addition to small sample equating conditions, the SiGNET equating design may also be preferred in other testing situations where pre-equating is desirable. For example, in many internet based testing (IBT) situations, it is desirable to have a new form pre-equated prior to administration, so that test takers can receive a scaled score immediately or soon after completing the test. This is probably one of the attractive features of IBT testing for test takers, because they can receive their scaled scores immediately after testing, compared to paper and pencil testing (PPT), where they often have to wait for several weeks before receiving a scaled score. Although the SiGNET equating design was used in the context of small samples in the current study, it can be readily adapted to IBT or other testing situations where pre-equating is needed.

Implications for Practice

There are some practical implications of using the SiGNET equating design rather than other equating designs. For example, using the SiGNET design requires administering more items than are operationally scored. This may not be feasible for some testing programs where the test taking time is already tight. Adding more items may introduce undesirable side effects such as test speededness. One way to address this issue is to increase testing time to allow more time for examinees to respond to the additional items. One can also evaluate whether increasing the number of items actually introduces test speededness. Finally, one can also make the

operational section of the test shorter if it does not considerably affect the current test reliability and content validity.

The SiGNET design seems like a viable option for most small volume MC tests, although for MC tests with fairly large number of subcontent categories, having a good representation of all the subcontent categories in the testlets may become difficult, especially when the number of items in each testlet is not very large. The SiGNET design may also not work very well with CR tests where the number of CR items in the test is quite small. For example, if a CR test has six CR items belonging to three content categories, it will be difficult to achieve content representation without creating a pretest testlet with at least one item from each content category, which may significantly increase the test taking time. Even if one were to include one CR item in the pretest testlet, using SiGNET may still result in a substantial increase in the test taking time, considering that the test had six items to begin with. Thus, for the SiGNET design to be successful, the testing program has to carefully evaluate tests on a case by case basis and decide whether using the SiGNET design would produce the added benefits without introducing other adverse consequences (e.g., increasing test speededness).

Finally, the SiGNET design assumes that the pretest items eventually will be scored in future new forms. However, there may be cases where all pretest items do not turn out to be good, and this may make the transition from the current form to a new form problematic. One way to address this issue would be to drop fewer items from the current form when making the transition to the new form. Consider a SiGNET equating situation where there are 20 pretest items, and therefore the plan is to exclude 20 currently operational items from the future new form (recall the drop testlet in Figure 2). If two pretest items perform poorly, then one way to make the transition to the new form is to drop 18 instead of 20 items, to account for the loss of the two pretest items.

Limitations and Future Research

A limitation of the current study is that the small samples in the SiGNET and the NEAT equatings were random samples drawn from one large population and are therefore quite equivalent in ability. Since previous research has shown that the SEE is somewhat larger when samples differ in ability (see Du Bose, 1993), it is reasonable to assume that in real testing situation where the samples from different testing administrations are often not equal in ability, the SEE would likely be larger than what was found in the current study. Therefore, future

research comparing SiGNET and NEAT equatings can systematically vary the ability of different samples and examine the effect on equating under the SiGNET and NEAT equating designs. For example, for the NEAT equatings, one can draw random samples for the new form that are lower in ability than the old form samples. Similarly one can draw samples of differing abilities for the different forms in the SiGNET equating chains.

A limitation of the SiGNET design is item exposure. Since all scored items in a new form under the SiGNET design are present in previous forms as pretest items, they are exposed more than they are in the NEAT design, where scored items in a new form are not necessarily present in previous forms. The precise effect of this exposure and its effect on SiGNET equatings cannot be easily determined. One way to examine this would be to draw samples under the SiGNET design where new form samples look progressively more able (as would be expected if items were overexposed) and examine the effect of such increasing examinee ability on the final equating results. Another way to examine this would be to formulate a model for how item exposure might affect test scores over repeated administrations, perhaps with the scores on each testlet-based form increasing with the number of times that form is given and also with the number of examinees who took that form. The performance of different equating approaches could be assessed with respect to this model.

In the SiGNET design, testlets are created to be mini versions of the total test and are nearly parallel in content and difficulty. When test developers create testlets, they may be more successful in creating mini versions that are nearly parallel in some situations than in other situations, just are they are often more successful in creating nearly parallel test forms in some situations and not so successful in other situations. Small differences in the testlets are not expected to create any problems in equatings, but the precise effect of large differences in a chain of SiGNET equatings is difficult to predict. Therefore, future studies may create testlets that have large differences in content and difficulty and evaluate the effect on SiGNET equatings. For example, Testlet 1 may have most items from one content area, such as reading, compared to Testlet 2, which may have most items from another content area, such as mathematics. Also, if statistics on these items are available, then one can deliberately create testlets of differing difficulty and examine the effect on testlet equating. One further area of research may be examining the impact on equating precision of factors such as the length of the

individual testlets (which has a direct relationship with the amount of overlap between new and old forms) and the number of testlets within each test.

Finally, this study used only the chained linear equating method to equate new test forms under the NEAT design. Although it is reasonable to expect that other linear equating methods would corroborate the results of this study (i.e., SiGNET equating resulting in a better equating in small sample conditions), future studies may use other linear equating methods such as Tucker, Levine, and Kernel (with large bandwidths) equatings to replicate the results of the current study. Finally, the current study used data from one basic skills test, and therefore future studies may use other tests and examine the effect on equatings under the SiGNET and NEAT designs. If results using other tests and equating methods are similar to the results of this study, the usefulness of the SiGNET design in small sample conditions and the minimum recommended sample sizes for equating under the NEAT and SiGNET equating designs can be more firmly established.

Despite these limitations and the fact that several issues were not addressed in this study, it seems reasonable to expect that the SiGNET equating design will prove to be very useful in situations where samples are just too small to conduct an equating under the NEAT design. To reiterate a point that was made earlier—-since proper data collection is regarded as the most important aspect of any equating—-the SiGNET equating design is an advantageous design in that it allows for a data collection design in which one can conduct an SG equating and also accumulate data to increase sample size when necessary.

References

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

Du Bose, P. (1993). *The accuracy and precision of linear equatings in small samples when examinee groups differ in ability*. Unpublished doctoral dissertation, University of South Florida, Tampa.

Grant, M. (2006, November). *Testlet design for equating with small samples: A means to equate small volume multiple-choice tests.* Paper presented at the Equating Special Interest Group Seminar, ETS, Princeton, NJ.

Hanson, B. A., Zeng, L., & Kolen, M. J. (1993). Standard errors of Levine linear equating. *Applied Psychological Measurement, 17,* 225–237.

Holland, P.W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.) *Educational measurement* (4th ed., p. 197). Westport, CT: Praeger Publishers.

Holland, P. W., Dorans, N. J., & Petersen, N. S. (2007). Equating test scores. In C.R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Volume 26. Psychometrics* (pp. 169–201). Amsterdam: Elsevier B.V.

Kim, S., von Davier, A. A., & Haberman, S. (2006). *Equating with small samples* (ETS Research Rep. No. RR-06-27). Princeton, NJ: ETS

Kolen, M. J. (1985). Standard errors of Tucker equating. *Applied Psychological Measurement, 9,* 209–223.

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

Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement,* 30, 23–29.

Parshall, C. G., Du Bose Houghton, P., & Kromrey, J. D. (1995). Equating error and statistical bias in small sample linear equating. *Journal of Educational Measurement, 32,* 37–54.

Petersen, N. S. (2007). Equating: Best practices and challenges to best practices. In N.J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 59–71). New York: Springer.

Skaggs, G. (2005). Accuracy of random groups equating with very small samples *Journal of Educational Measurement*, *42*(4), 309–330.

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

Notes

[1] Although with the SiGNET design, any set of items can be used instead of testlets, using testlets is preferred because it is easier to maintain the content balance of the test when using testlets. For example, if a test consists of reading, writing, and mathematics subsections and the additional unscored testlet consists of a few items from all three subsections, then adding that unscored testlet does not disrupt the content balance of the test. However, if the unscored testlet consisted of items from one subsection only, then the full test might appear to be heavily weighted towards that subcontent area. Other advantages of using testlets as opposed to any random item set are given in Grant (2006).

[2] Under the NEAT design, a chained linear equating was used to equate the scores of Form 5 to scores on Form 1. With this procedure, new form scores are converted to scores on the common items using results from examinees who took the new form. Next, scores on the common items are equated to scores on the old form using results from examinees who took the old form. These two conversions are then chained together to produce a conversion of the new form scores to the old form scores (Kolen & Brennan, 2004).

[3] Although not initially planned, it seemed worthwhile to extend the analyses to examine what sample sizes would be required to have a NEAT equating under this condition (where the new form has about 25% overlap with the old form) that would be trustworthy (in terms of lower sampling error and bias). Several NEAT equatings were run with different sample sizes such as 100, 200, 300, 400, 500, 600, and 700, to evaluate what minimum sample size is required to conduct a trustworthy NEAT equating (i.e., equating Form 5 directly to Form 1). Results indicated that the average SEEs and average RMSD were both 0.528, which is close to the DTM criterion. The bias value was close to zero for these sample size conditions. Therefore, for testing programs more interested in average SEE and bias, these sample sizes may be acceptable for conducting chained linear equating under the NEAT design. But for tests that employ cut scores, the conditional SEEs may be more informative than the overall error. As seen in Figure A2, the CSEEs were larger than 0.5 for several score points in the 35 to 61 score range for the 300 sample size condition. As seen in this graph, the CSEEs were lower than 0.5 for the 35 to 61 score region when the sample size was at least 600. For sample sizes of 600, the conditional bias was also close to zero for all score points. Therefore for this test, one would need at least a sample size of 600 to ensure that equated scores (using chained

linear equating) within the ±1.5 *SD* range from the mean are trustworthy (especially in terms of lower CSEEs). The word *test* is italicized to emphasize that these results were true for this test. However, with other tests where different conditions may apply (i.e., anchor versus total correlation may be different, the scores may be more evenly distributed throughout the score scale, etc), one may obtain somewhat different results.

[4] Although a DTM criterion of 0.5 was used in the current study, other criteria have also been widely used to evaluate equating precision. One such criterion suggested by Kolen and Brennan (2004, pp.258–261) is the 1/10th of the standard deviation unit observed in the total sample. CSEEs larger than this value are considered large. If this approach were used in the current study, then the criterion value would be 0.883 (0.1 multiplied by 8.83, which is the *SD* of Form 5 in the total sample). Since this value is larger than the DTM value of 0.5, fewer examinees would be required to achieve equating precision relative to this new criterion.

Appendix

Conditional Standard Error of Equating for Single Group Nearly Identical Tests (SiGNET)

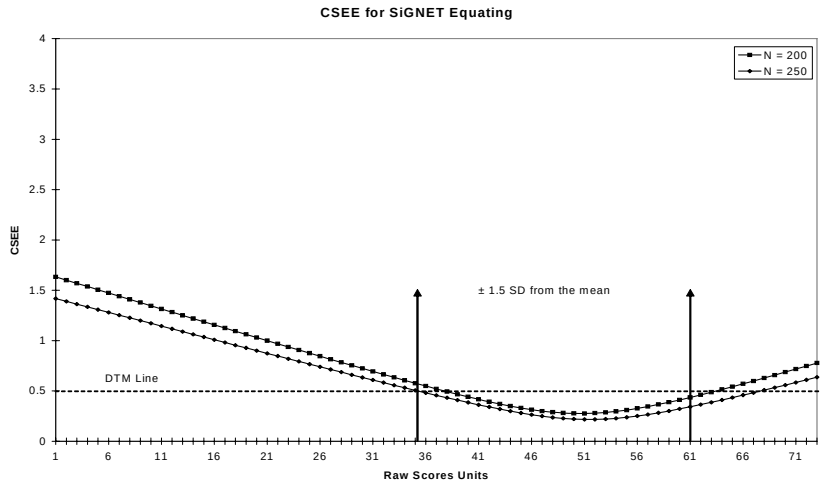

Figure A1. **Conditional standard error of equating for single group nearly identical tests (SiGNET, Form 5 to Form 1 equatings) for sample sizes of 200 and 250.**

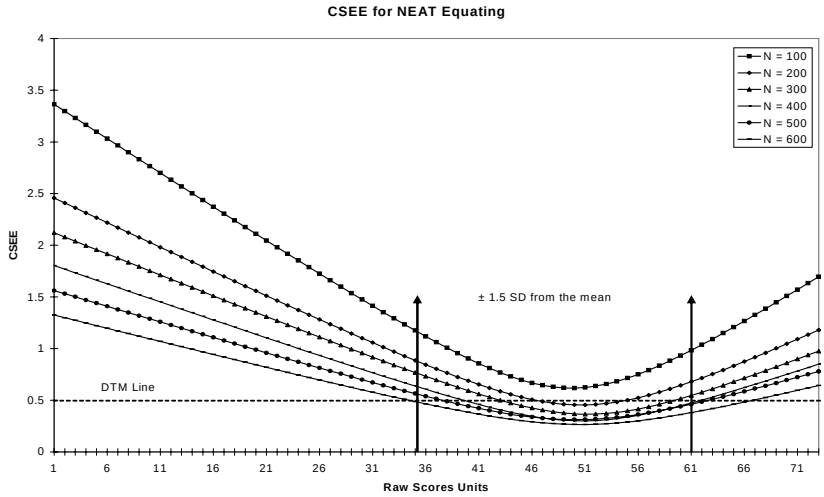

Figure A2. **Conditional standard error of equating for the nonequivalent groups with anchor test (NEAT) design (Form 5 to Form 1 equatings) for sample sizes 100, 200, 300, 400, 500, and 600.**