Listening. Learning. Leading.®

Research Report
ETS RR-13-02

# Exploring Alternative Test Form Linking Designs With Modified Equating Sample Size and Anchor Test Length

Lin Wang

Jiahe Qian

Yi-Hsuan Lee

February 2013

# ETS Research Report Series

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

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

# Exploring Alternative Test Form Linking Designs With Modified Equating Sample Size and Anchor Test Length

Lin Wang, Jiahe Qian, and Yi-Hsuan Lee
ETS, Princeton, New Jersey

February 2013

**Action Editor:** Rebecca Zwick

**Reviewers:** Anna Kubiak and Yue Jia

.

# Abstract

The purpose of this study was to evaluate the combined effects of reduced equating sample size and shortened anchor test length on item response theory (IRT)-based linking and equating results. Data from two independent operational forms of a large-scale testing program were used to establish the baseline results for evaluating the results from two alternative designs. Under the two alternative designs, two simulated conditions were created from the original data. Under one condition, we reduced the equating sample size (from about 2,000 to about 1,000) per anchor item and shortened the anchor test length (by half) per equating sample. Under the other condition, we reduced the sample size (from about 2,000 to about 1,000) per anchor item only. A complete grouped jackknife replication method was used to estimate the standard errors of the linking and equating procedures from 100 jackknife replicate samples; the complete procedures included IRT calibrations, item parameter scaling, and IRT true score equating. The findings from a comparison of the results from the two simulated conditions and the baseline results showed that neither alternative design had any practical impact on the linking and equating results for either test form.

Key words: anchor test length, equating sample size, equating to a calibrated item pool design, complete grouped jackknifing

## Acknowledgments

This paper was presented at the 2011 annual meeting of National Council on Measurement in Education (NCME) on April 10, 2011, New Orleans, Louisiana.

## 1. Overview

For large-scale testing programs that administer multiple similar forms over time, scores from different test forms must be interchangeable, according to the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). In addition to test development efforts to produce forms of similar construct representation and difficulty, test form equating is a required psychometric procedure for reporting comparable scores. A variety of data collection designs and equating methods have been developed over the past few decades to equate tests (Holland, 2007).

One frequently used data collection design is a nonequivalent group with anchor test (NEAT) design. Under this design, a set of common items (called anchor items, or anchors for short) is delivered in two administrations that are to be linked (we assume one distinct test form per test administration in this report). Common items are used because the two test forms might have small differences in their difficulty level and/or because the examinees in the two administrations might differ in their ability levels as measured by the test. A group ability difference may be confounded with the test form difficulty in the equating process, which is designed to adjust for minor differences in difficulty between a new test form and its reference form. Therefore, it is necessary to separate the group ability difference from the test form difficulty difference during equating. A testing program may choose to use internal anchors (anchor items contributing to test scores), external anchors (anchor items not contributing to test scores), or both. Both item response theory (IRT)-based methods and classical test theory methods can be employed in equating. If IRT methods are used, equating also includes the linking of item parameters from different administrations.

In a true NEAT design, a new form may be equated to one or more reference forms through carefully selected anchor items. In IRT-based equating applications, a design called *common item equating to a calibrated item pool* (Kolen & Brennan, 2004, pp. 201-205) can be implemented (*equating to a pool* for short). Like the NEAT design, the equating to a pool design also uses anchor items. However, this design differs from NEAT in that it is not necessarily a form-to-form equating, but a new-to-base form equating through the calibrated items. A base form is the test form on which the IRT scale for the testing program is established. The raw-to-scale-score conversion table on the base form is used to convert equated raw scores on a new form to scaled scores. Specifically, under the equating to a pool design, anchor items on a new form do not have to come from the same reference form. Instead, anchor items may be selected according to specified rules from an item pool that consists of scaled items (items whose IRT parameters are all on the IRT scale of the base form) from all previously administered test forms. The anchor items are then administered along with the new items on a new form. The raw scores on the new form are subsequently equated to the base form instead of being equated to any particular reference form(s) from which the anchor items are taken. After the new form is equated, the new items are put on the IRT scale of the test.

In both the NEAT and the equating to a pool designs, it is reasonable to assume that, in general, a longer anchor test (more anchor items) would help with the quality of the linked scores and that larger equating samples (more examinees taking the same anchor items) would produce more stable results. Anchor items' content representation in the total test is often considered a necessity such that the anchor set functions as a minitest. With a longer anchor test, it is more likely that the anchor will resemble the test in terms of content. This in turn may increase the correlation between the anchor and the test, which improves the quality of equating.

The subject of sample and anchor sizes in equating has received a fair amount of treatment in the educational measurement literature. Guidelines have been suggested for practitioners to determine appropriate sample sizes and anchor sizes in real world testing situations (Kolen & Brennan, 2004). Studies have been conducted to explore how equating results would be impacted by such factors as anchor sizes (Klein & Kolen, 1985; Ricker & von Davier, 2007), anchor items' content representation of the test (Klein & Jarjoura, 1985), or the combined effects of anchor size, content representation, and some other factors, such as anchor difficulty relative to test difficulty, as well as the distribution of anchor item difficulty (Liao, 2009; Sinharay & Holland, 2007).

What prompted this study was a unique situation that had developed in a large testing program and that was not addressed in the existing literature. In this situation, the availability of typical equating data with respect to sample sizes, anchor test length, and the distribution of anchor items in equating samples (number of anchor items seen by each examinee in an equating sample) would be limited. Specifically, new operational testing conditions were being proposed that might not allow for the administration of all anchor items to the same group of examinees as was the current practice, but would  require splitting the anchor items evenly between two groups of examinees so that each examinee’s workload could be reduced. Therefore, two equating samples would be needed for the new situation, compared to only one equating sample in the current practice. In addition, the number of examinees that would see each anchor item would have to be reduced in order to accommodate the first change (splitting the anchor items between two groups), because testing volume (number of examinees available for each test administration) is another limiting factor in operational administrations. With these two changes, the existing anchor length would be shortened per equating sample so that each examinee in an equating sample would see half of the total anchor items for a new form; the equating sample sizes per anchor item would be reduced so that fewer examinees would see each anchor item (these will be the definitions of the reduced equating samples and the shortened anchors in this report). Because both the equating sample size and the anchor test length contribute to the results of IRT calibration, linking, and equating, such a change should not be made without evaluating its possible impact on linking and equating results.

This study was carried out to accomplish two objectives: One was to evaluate the joint effects of the above-mentioned two changes on the results of IRT linking and equating; the other was to evaluate the impact on the linking and equating results from reduced sample size alone. The following research questions guided this study and the interpretation of the findings:

How would the reduced sample size and the shortened anchor test length jointly affect the IRT linking and equating results?

How would the reduced sample size alone affect the IRT linking and equating results?

Would the results be replicable with new data?

Section 2 is the method section that describes the data source, the designs for this study, and the planned analyses. Results of the study are presented in section 3, followed by a discussion and the conclusion in section 4.

## 2. Method

### 2.1 Data Source

This study used existing data from two independent operational administrations of a large-scale testing program. The same data collection designs and analyses were applied to data from each administration. Specifically, each administration delivered one new test form that contained both operational and anchor test items. The two new forms (called Form 1 and Form

2) had no overlapping items and were administered about 6 months apart. These two new forms were selected for this study because about 2,000 examinees were available for the anchor items on each new form, so that it would be possible to split the examinees into two groups of about 1,000 each for the purpose of this study. The overall sample sizes for Form 1 and Form 2 were 8,010 and 7,360, respectively. The equating design for the operational administrations used both internal and external anchor items. Every examinee who took either new form saw all the internal anchor items on that test form, because those items were part of the operational test. The external anchor items on each new form were seen by about 2,000 examinees. This study focused on the external anchors, because these were the ones that could be shortened for the study without changing the length of the operational test; the internal anchor items were simply considered operational items and were not mentioned separately in this method section. The focus on the external anchors in this study was based on two considerations. First, internal anchors have been and will continue to be used together with external anchors in operational equating. The design of this study followed this practice for operational equating. Second, the effect of the internal anchors, which counted for half of the total anchors, was considered constant because everyone saw the internal anchors; any differences in the equating results from this study would be due to the changes to the external anchors.

All the operational and external anchor items were set-based; each set had one stimulus and 14 questions (items). Each form (Form 1, Form 2) contained three operational sets and two external anchor sets. Each set of the external anchor items can be considered a subanchor and a minitest of the total test in terms of the content representation. The statistical properties of the operational and anchor sets on each new form were controlled to be relatively close to one another. Tables 1 and 2 provide the descriptive statistics of the items on Forms 1 and 2, respectively, including the classical item statistics, such as average difficulty level (*P*+) and biserial correlation (Biserial) for item discrimination, and two-parameter logistic (2PL) IRT model-based item parameters, such as discrimination parameter *a* and difficulty parameter *b* (Lord, 1980; Lord & Novick, 1968). The anchor items in the two tables refer to the external anchor items. The correlations between the anchor tests and operational tests were 0.84 and 0.83 for Form 1 and Form 2, respectively. On Form 1, the anchor test was easier than the operational test in terms of IRT parameter *b* values (-0.81 vs. -0.65). On Form 2, however, the two were very close (-0.58 vs. -0.56).

**Table 1**

***Descriptive Statistics of the Anchor and the Test Items on Form 1***

| Form 1 | Item | Classical item statistics | | | | 2PL IRT item parameter estimates | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | P+ | | Biserial | | *a* | | *b* | |
| | | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| All anchor items | 26 | 0.72 | 0.14 | 0.51 | 0.08 | 0.70 | 0.23 | -0.81 | 0.78 |
| Subanchor 1 | 13 | 0.70 | 0.10 | 0.53 | 0.07 | 0.71 | 0.23 | -0.71 | 0.47 |
| Subanchor 2 | 13 | 0.74 | 0.17 | 0.49 | 0.09 | 0.68 | 0.23 | -0.90 | 1.01 |
| Operational items | 42 | 0.67 | 0.15 | 0.52 | 0.08 | 0.59 | 0.20 | -0.65 | 0.96 |

*Note*. *N* = 8,010. 2PL = two-parameter logistic, IRT = item response theory.

**Table 2**

***Descriptive Statistics of the Anchor and the Test Items on Form 2***

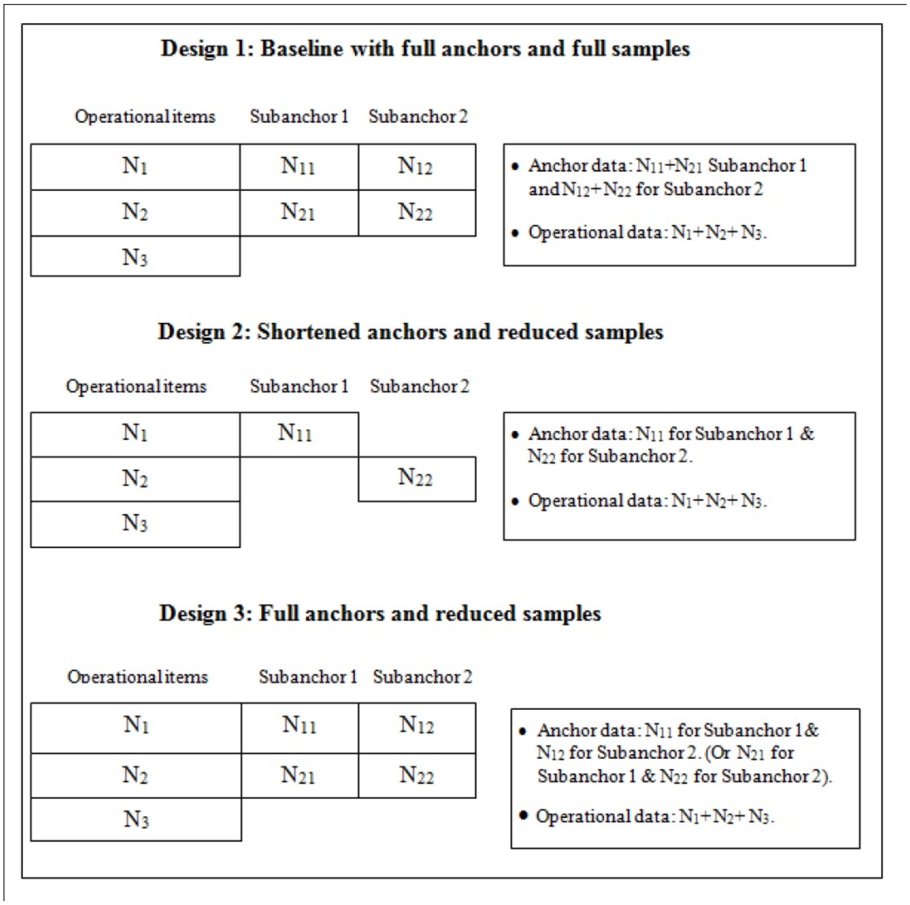
| Form 2 | Item | Classical item statistics | | | | 2PL IRT item parameter estimates | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | P+ | | Biserial | | *a* | | *b* | |
| | | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| All anchor items | 27 | 0.62 | 0.12 | 0.45 | 0.09 | 0.59 | 0.18 | -0.58 | 0.68 |
| Subanchor 1 | 13 | 0.64 | 0.13 | 0.42 | 0.08 | 0.55 | 0.17 | -0.69 | 0.78 |
| Subanchor 2 | 14 | 0.61 | 0.12 | 0.47 | 0.09 | 0.64 | 0.20 | -0.48 | 0.58 |
| Operational items | 42 | 0.63 | 0.14 | 0.53 | 0.08 | 0.64 | 0.19 | -0.56 | 0.77 |

*Note*. *N* = 7,360. 2PL = two-parameter logistic, IRT = item response theory.

Based on the type of analysis to be described in this section, the operational data from each form were first analyzed to establish a baseline. The operational data were then used to create the two conditions of reduced-sample and shortened-test anchors and to build jackknife replicate samples. The data on each operational form contained all examinees' responses to all the operational items and about 2,000 examinees' responses to the external anchor items on that form. For convenience of discussion, the examinees that were administered the external anchor items will be referred to as *equating samples*. In Tables 1 and 2, *All anchor items* refers to the external anchors; *Subanchor 1* and *Subanchor 2* refer to the two external anchor sets.

## 2.2 Three Designs of Anchor Data Structures

Figure 1 illustrates the three designs for this study. Design 1 shows the *baseline condition*. Under this design, N1 and N2 refer to the data collected from all examinees that saw the full anchor (all anchor items). In $N_{11}$, $N_{12}$, $N_{21}$, and $N_{22}$, the first subscript refers to the equating sample; the second subscript indicates each subanchor (Subanchor 1 and Subanchor 2). For example, $N_{11}$ is for the data collected from the first equating sample of about 1,000 examinees that saw the first subanchor items (Subanchor 1). In this design, all the examinees' responses to all the operational and anchor items were used for IRT item calibration. The responses to all the anchor items from about 2,000 examinees (about 1,000 from $N_{11}$ and $N_{21}$, and about 1,000 from $N_{12}$ and $N_{22}$) were used in the linking and equating procedures. N3 consisted of the examinees that saw only the operational items; the sample size of N3 was the total sample minus the equating sample. N3 remained the same for all three designs.



**Design 1: Baseline with full anchors and full samples**

| Operational items | Subanchor 1 | Subanchor 2 |
|---|---|---|
| $N_1$ | $N_{11}$ | $N_{12}$ |
| $N_2$ | $N_{21}$ | $N_{22}$ |
| $N_3$ | | |

- Anchor data: $N_{11}$+$N_{21}$ Subanchor 1 and $N_{12}$+$N_{22}$ for Subanchor 2
- Operational data: $N_1$+$N_2$+$N_3$.

**Design 2: Shortened anchors and reduced samples**

| Operational items | Subanchor 1 | Subanchor 2 |
|---|---|---|
| $N_1$ | $N_{11}$ | |
| $N_2$ | | $N_{22}$ |
| $N_3$ | | |

- Anchor data: $N_{11}$ for Subanchor 1 & $N_{22}$ for Subanchor 2.
- Operational data: $N_1$+$N_2$+$N_3$.

**Design 3: Full anchors and reduced samples**

| Oɒerational items | Subanchor 1 | Subanchor 2 |
|---|---|---|
| $N_1$ | $N_{11}$ | $N_{12}$ |
| $N_2$ | $N_{21}$ | $N_{22}$ |
| $N_3$ | | |

- Anchor data: $N_{11}$ for Subanchor 1& $N_{12}$ for Subanchor 2. (Or $N_{21}$ for Subanchor 1 & $N_{22}$ for Subanchor 2).
- Operational data: $N_1$+$N_2$+$N_3$.

***Figure 1.*** **Three designs of anchor data structures.**

Design 2 depicts the new condition of the reduced equating sample and the shortened anchor test length. From the operational data of Design 1, a new data set was created to contain a randomly selected sample of about 1,000 examinees (half of the 2,000) who responded to the items on one of the two subanchors (Subanchor 1 or Subanchor 2). The selection of a subanchor was also random. As a result, for the N1 group, the 1,000 examinees' responses to Subanchor 1 ($N_{11}$) were retained, but their responses to Subanchor 2 ($N_{12}$) were excluded. Similarly, for the N2 group, the 1,000 examinees' responses to Subanchor 2 ($N_{22}$) were retained, and their responses to Subanchor 1 ($N_{21}$) were excluded. The retained data from the N1 group on Subanchor 1 ($N_{11}$) and from the N2 group on Subanchor 2 ($N_{22}$) were combined with the N3 group for the analyses in Design 2.

Although the examinees in $N_{11}$ and $N_{22}$ under Design 2 saw only one set of anchor items, all the examinees ($N_{11}$, $N_{22}$, N3) saw the same operational items. Therefore, one IRT calibration of all the items (operational and external anchor) would put their item parameter estimates on the same scale (Kim & Cohen, 1998; Lord, 1980). The IRT parameter estimates of the items were next transformed (scaled) to the IRT base scale of the test. Although Design 2 bears some resemblance to the double-linking practice under a NEAT equating design, in which a new form is equated to two reference forms using one set of common items from each reference form, there is a difference between Design 2 and the double-linking design. Under a double-linking design, each anchor test typically contains the full anchor containing the required number of items for equating. For example, a testing program may require using 30 common items to link a 100-item new form to its reference form. On occasion, however, the program might add another set of 30 common items to link the same new form to a second reference form to enhance the equating quality. The new form is then equated to each reference form; the results from the two equatings (one per reference form) would typically be combined in an appropriate manner for the final equating results of the new form. Under Design 2, however, only one subanchor (half of the full anchor) was seen by each equating sample, which was also half the size of the original equating sample.

Design 3 portrays the effect of reduced equating sample size only; it differs from Design 2, in which the effect of the reduced sample sizes was confounded with that of the shortened anchor test length. In Design 3, the full equating sample of about 2,000 was divided into two random samples of about 1,000 each; for each equating sample, the examinees' responses to the full anchors were retained for analysis.

For each of the three designs above, 100 jackknife replicate samples were constructed as follows:

1. The examinees in the sample were sorted randomly and aggregated into groups of similar sizes. When *N* was not the multiple of 100 (that is true for most situations), there were some cases left (*N'* < 100). One case was selected from the remaining cases and assigned to the first group; then another case from those remaining was selected and assigned to the second group, and so on until no cases were left. Therefore, 100 jackknife replicate samples were formed in total. For example, in this study data, Form 1 had 8,010 examinees. The 8,010 examinees were randomly assigned to 100 disjoint groups of 80 per group. The 10 remaining examinees were then randomly assigned to the first 10 jackknife groups, one per group, in the way described above.
2. Let *J* be the total number of groups formed, *J* = 100 in this study*;* then, the *j*th jackknife replicate sample was formed by deleting the *j*th group (*n* = 80 or 81) from the whole sample (*N* = 8,010). Therefore, each replicate sample size was either 7,930 or 7,931.
3. The item response data in each of the 100 replicate samples were used for IRT calibration, scaling, and equating. This procedure is called a *complete grouped jackknife repeated replication* (CGJRR) method. The CGJRR procedure, programmed in SAS, included the step of IRT calibration and the steps of parameter scaling and IRT true score equating.

### 2.3 Analysis

The following analyses were conducted:

1. The linking and equating of the new form proceeded in three steps: IRT calibration, item parameter transformation (scaling), and IRT true score equating. This was carried out on each of the 100 jackknife samples under each of the three designs. A 2PL IRT model was chosen for item calibration using the Parscale software package (ETS version by Muraki & Bock, 1999). The test characteristic curve (TCC) method by Stocking and Lord (1983) was used to obtain the scaling coefficients *A* and *B* from

the anchor items. The scaled parameter estimates of the operational items of the new form were then used in the IRT true score equating step, in which the expected raw scores on the new form were equated to the raw scores of the test's base form (not the reference forms from which the common items were taken). Both the item parameter scaling and the IRT true score equating steps were implemented using the ICEDOG software developed at ETS (Robin, Holland, & Hemat, 2006).

2. The statistics of interest in this study were the scaling coefficients A and B (note, not item discrimination parameter a, or difficulty parameter b), the equated raw scores, the sample means of scaled scores, and sampling error estimates. The aforementioned CGJRR procedure was used to estimate the sampling errors of the whole linking and equating procedures (Qian, 2005). For the whole sample and each jackknife replicate sample, the same IRT calibration, scaling, and equating procedures were carried out, and then the jackknifed standard errors of the parameters of interest were estimated. Let $\hat{\theta}$ be the parameter estimated from the whole sample and $\hat{\theta}_{(j)}$ be the estimate from the jth jackknife replicate sample. The jackknifed variance of $\hat{\theta}$ was estimated by

$$v_J\left(\hat{\theta}\right)=\frac{J-1}{J}\sum_{j=1}^{J}\left(\hat{\theta}_{(j)}-\hat{\theta}\right)^2 ,$$

and the jackknifed standard error was then computed by

$$se_J\left(\hat{\theta}\right)=\sqrt{v_J\left(\hat{\theta}\right)}$$

(Wolter, 2007). Formula $v_J\left(\hat{\theta}\right)$ has alternative forms. For example, $\hat{\theta}$ can be replaced by the mean of all $\hat{\theta}_{(j)}$ (*j* = 1, 2, …, *J*).

3. The aggregated results from the three designs were compared to evaluate the effects of the reduced sample size and the shortened anchor test length.

## 3. Results

Table 3 summarizes the results for Form 1. The means and jackknifed standard errors are presented in the top section of Table 3. As was stated before, the results of Design 1 were considered the baseline results to be compared with the results of Design 2 and Design 3, respectively. In the Difference section of Table 3, the differences in the average scaling constants *A* and *B*, and the averaged scaled scores between Design 2 and Design 1 and between Design 3 and Design 1 are presented, respectively. The comparisons of interest (means of *A*, means of *B*, and means of mean scaled score) all showed small differences; none of the differences was statistically significant (t-test, $p < 0.05$). For example, the differences in the scaling constants *A* and *B* were 0.0048 and 0.0092, respectively, between Design 2 (shortened anchors, reduced samples) and Design 1 (full anchors, full samples); neither of the differences was statistically significant. The ratios of the jackknifed standard errors (jackknifed SE in the bottom section of Table 2) were close to (but greater than) 1 in both comparisons. Although the ratios of the jackknifed standard errors were greater than 1, the jackknifed standard error values were too small to be of any practical significance in all cases. Design 1 had the smallest jackknifed standard error among the three designs for any statistic of interest. This was expected, because Design 1 used the full anchors and full samples.

**Table 3**

***Summary Statistics From the Jackknifed Samples on Form 1 for the Three Designs***

| Form 1 data | Mean scaling constant A | Jackknifed SE(A) | Mean scaling constant B | Jackknifed SE(B) | Mean scaled score | Jackknifed SE (mean) |
|---|---|---|---|---|---|---|
| Design 1 (full anchors, full samples) | 1.1209 | 0.0143 | 0.2999 | 0.0158 | 20.88 | 0.080 |
| Design 2 (shortened anchors, reduced samples) | 1.1257 | 0.0162 | 0.3091 | 0.0169 | 20.92 | 0.087 |
| Design 3 (full anchors, reduced samples) | 1.1163 | 0.0175 | 0.2909 | 0.0177 | 20.84 | 0.092 |
| Design 2 – Design 1 | 0.0048 | - | 0.0092 | - | 0.040 | - |
| Design 3 – Design 1 | –0.0046 | - | –0.0090 | - | –0.040 | - |
| Design 2 / Design 1 | - | 1.1330 | - | 1.0700 | - | 1.088 |
| Design 3 / Design 1 | - | 1.2240 | - | 1.1200 | - | 1.148 |

Table 4 summarizes the results for Form 2. The differences from the comparisons of interest are shown in the Difference section of Table 4. As in Table 3, by t-test, none of the comparisons of interest showed statistically significant differences. Similarly, the differences were all very small. The ratios of the jackknifed standard errors for the compared designs were all close to 1; the actual values of the errors were too small to be of any concern in practice.

**Table 4**

***Summary Statistics From the Jackknifed Samples on Form 2 for the Three Designs***

| Form 2 data | Mean scaling constant A | Jackknifed SE(A) | Mean scaling constant B | Jackknifed SE(B) | Mean scaled score | Jackknifed SE(mean) |
|---|---|---|---|---|---|---|
| Design 1 (full anchors, full samples) | 1.0317 | 0.0135 | 0.0846 | 0.0156 | 19.76 | 0.096 |
| Design 2 (shortened anchors, reduced samples) | 1.0222 | 0.0143 | 0.0817 | 0.0170 | 19.78 | 0.105 |
| Design 3 (reduced samples, full anchors) | 1.0406 | 0.0159 | 0.0914 | 0.0176 | 19.78 | 0.105 |
| Design 2 – Design 1 | – 0.0095 | - | – 0.0029 | - | 0.01 | - |
| Design 3 – Design 1 | 0.0089 | - | 0.0068 | - | 0.02 | - |
| Design 2 / Design 1 | - | 1.0593 | - | 1.0859 | - | 1.092 |
| Design 3 / Design 1 | - | 1.1771 | - | 1.1217 | - | 1.088 |

Figure 2 depicts the differences in the equated raw scores between the designs for Form1 and Form 2, respectively. Although the equated raw score differences between the two designs differed in their patterns, the differences in the unit of the raw score were small for all designs on both forms. For example, the largest difference on Form 1 was less than 0.1 raw score points between Design 2 and Design 1. On Form 2, the largest difference was slightly over 0.1 raw score points between Design 2 and Design 1. For the test used, the raw score scale ranged from 0 to 42 and the reporting scale ranged from 0 to 30, both in units of 1. This 0.1 point difference on the raw score scale would translate into a difference of slightly over 0.1 but less than 0.5 points on the reporting scale. Using the criterion of the difference that matters (DTM) recommended first by Dorans and Feigenbaum (1994) and reiterated by Holland and Dorans (2006, p. 212), the

DTM for this test would be 0.5 points on the reporting score scale. Therefore, the small difference in the equated raw scores between Designs 1 and 2 would not have any practical consequence to the scores that would be reported to the examinees.

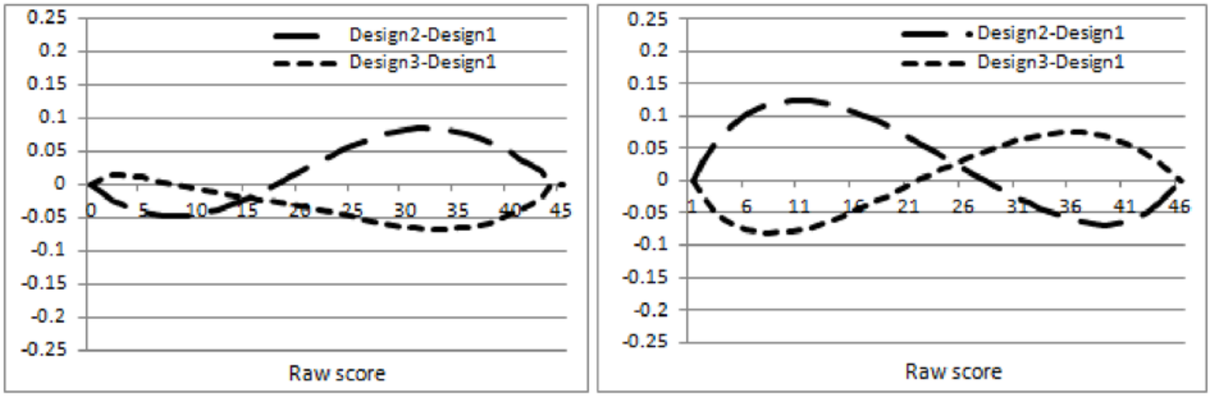


***Figure 2*. Equated raw score differences between the designs (left Form 1, right Form 2).**

Figure 3 displays the comparison of the jackknifed standard errors (JRR SE) across all the equated raw scores among the three designs on both forms. These were conditional standard errors. All the JRR SE values were small: less than 0.18 on Form 1 (left chart), and less than 0.20 on Form 2 (right chart). The baseline (Design 1) showed the smallest SE values on both forms, followed by Design 2, and then Design 3.

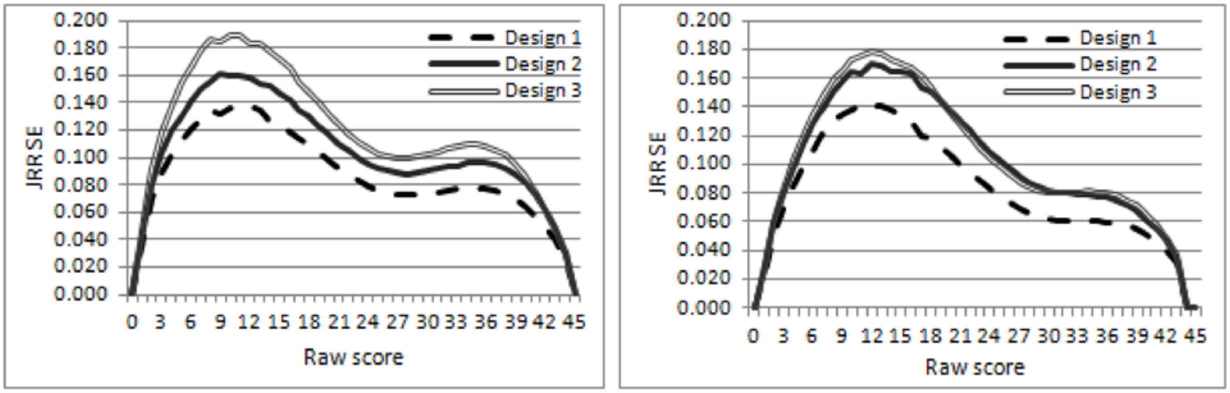


***Figure 3*. Jackknifed standard errors from the three designs (left Form 1, right Form 2).**

# 4. Discussion and Conclusion

## 4.1 Discussion

This study used existing operational data to simulate two situations. In one simulation, the equating sample sizes were reduced from about 2,000 (Design 1) to about 1,000 (Design 3) with every candidate seeing all the anchor items (full anchor). In the other situation, the full anchor was divided into two subanchors (Subanchor 1, Subanchor 2) of equal size, and so was the original equating sample (N1, N2). Although each examinee saw the full anchor, for each new equating sample, only the examinees' responses to the items of one subanchor (in $N_{11}$, $N_{22}$, under Design 2 in Figure 1) were used for calibration, linking, and equating. This study was designed to explore how the reduced sample size and the shortened anchor test length jointly would affect the IRT linking and equating results, how the reduced sample size alone would affect the IRT linking and equating results, and whether the results would be replicable with new data. Under the equating to a calibrated item pool design for this study, both equating sample size and anchor test length were critical factors that would impact the quality of equating.

The results for Design 2 would answer the first research question. Design 2 yielded small differences from Design 1 in both the means and jackknifed standard errors on the variables of interest: the scaling constants *A* and *B,* and the equated raw scores. However, all the differences were too small to be of any practical significance or impact. In other words, Design 2 did not appear to show practical differences from the baseline of Design 1 in either the IRT parameter linking results (scaling constants *A* and *B*), or the equating results (equated raw scores). This would suggest that, at least for this study with the given characteristics of the anchor items and the equating samples, Design 2 could replace Design 1 as the equating data collection design in operational work; this change would lead to little practical impact on equating results. This finding is not totally unexpected. Although the full anchor was split into two subanchors so that each subanchor (one set) was seen by only about 1,000 instead of 2,000 examinees, and so that each examinee would see one subanchor instead of the full anchor (two sets), the total anchor test length was not shortened for the entire form (Form 1 or Form 2), and each form was still equated using the full anchors. It is apparent that using the two subanchors would yield very small differences in the results compared to using the results of the full anchor that was seen by the full equating sample.

The small differences in the results between Design 3 and Design 1 indicated that the equating sample size for such an application could be reduced to about 1,000 with little practical impact. This is the answer to the second research question. The findings from this study are important in that very small differences can be expected in the linking and equating results between equating sample sizes of 2,000 and 1,000.

Form 2 was included in this study to check whether the results on Form 1 would be replicated when the identical designs and analyses used on Form 1 data were used in analyzing the Form 2 data. The results from the analyses of Form 2 also showed small linking and equating differences between Design 2 and Design 1, and between Design 3 and Design 1. Because the data from Form 2 did not overlap with the data from Form 1, the similarity of the findings from the data on the two forms added confidence to the expectation that Design 2 and Design 3 would produce similar results for new test forms.

Designs 2 and 3 and the CGJRR jackknife method may be employed in practice to evaluate impact due to reduced equating sample sizes and shortened anchor test lengths. The findings from this particular study were encouraging, as they appeared to support Design 2 in practice for efficient use of sample size and anchor test length by tests that may resemble the test for this study. To improve test access to examinees so that they have more opportunities to sign up for a test, a more frequent test administration schedule may be needed. This might mean that a smaller sample is available in each test administration, which would require more efficient use of the available data for required psychometric analyses such as IRT calibration and equating. Findings from this study could be valuable to a testing program that aims at both maintaining the psychometric quality of the tests and sustaining a high frequency of test administrations.

### 4.2 Limitations of the Study and Suggestions for Future Research

Design 2 was the main focus of the study for assessing the combined effects of a reduced sample size per item and a shortened anchor test length per examinee. Under this design, each simulated equating sample was formed by randomly assigning an examinee to one of the two simulated equating samples and then selecting this examinee’s responses to only the items of one subanchor. This design led to two characteristics of the two simulated equating samples. One characteristic was that, as a result of the random selection, the two samples were assumed to have similar ability levels in their performance on the whole test. The other characteristic was that, in each simulated equating sample, each examinee’s responses to the items of only one anchor set (half of the full anchor) were used in all the analyses. These two characteristics resulted in two limitations of this study. First, the similar group ability levels of the two simulated equating samples did not offer the study an opportunity to investigate what would happen if the two simulated equating samples (formed on the two subanchors) would have differed in their ability levels. In operational testing, it may not always be possible to obtain equating samples of similar ability; the ability levels of two equating samples may differ due to the effects of seasonality (when the test is administered), nonrandom delivery of anchor items to equating samples, missing data, and so forth. Differences in ability between equating samples may affect linking and equating results. Second, using the data from only one of the two subanchors might be subject to possible effects induced by some dynamics between a particular equating sample and a particular subanchor (some degree of sample by anchor interaction). As a result, the findings from this study cannot be readily generalized to other equating situations that use different data collection designs.

It is suggested that future research along these lines consider these two limitations and incorporate designs that would simulate conditions of varying group ability levels when creating new equating samples from the original sample (considered as the population). Similarly, for each equating sample under Design 2, the data on both subanchors could be analyzed separately. For example, under Design 2, only the item data on Subanchor 1 ($N_{11}$) from the first equating sample (N1) and the item data on Subanchor 2 ($N_{22}$) from the second equating sample (N2) were combined for linking and equating in this study. In a future study, the item data on Subanchor 2 ($N_{12}$) from the first equating sample (N1) and the data on Subanchor 1 ($N_{21}$) from the second equating sample (N2) may be combined for linking and equating. This would provide the opportunity to evaluate jackknifed errors from the perspective of within-person random selections of item responses.

# References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (ETS Report Memorandum No. RM-94-10; pp. 91–122). Princeton, NJ: Educational Testing Service.

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

Holland, P. W. (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 5–30). New York, NY: Springer.

Kim, S. H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under the item response theory. *Applied Psychological Measurement*, *22*, 131–143.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement*, *22*(3), 197–206.

Klein, L. W., & Kolen, M. J. (1985, April). *Effect of number of common items in common-item equating with nonrandom groups*. Paper presented at the annual meeting of American Educational Research Association, Chicago, IL.

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

Liao, C. (2009, April). *Examining the impact of anchor test selection in a sub-optimal equating situation for pseudo-tests created from a large-scale English proficiency test*. Paper presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, San Diego, CA.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Erlbaum.

Lord, F. M., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Muraki, E., & Bock, R. (1999). *PARSCALE 3.5: IRT item analysis and test scoring for rating-scale data*. Lincolnwood, IL: Scientific Software, Inc.

Qian, J. (2005, April). *Measuring the cumulative linking errors of NAEP trend assessments.* Presentation at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Ricker, K., & von Davier, A. (2007). *The impact of anchor test length on equating results in a nonequivalent groups design* (ETS Research Report No. RR-07-44). Princeton, NJ: ETS.

Robin, F., Holland, P., & Hemat, L. (2006). ICEDOG [Computer software]. Princeton, NJ: ETS.

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44*(3), 247–275.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210.

Wolter, K. (2007). *Introduction to variance estimation* (2nd ed.). New York, NY: Springer.