LaFlair, G. T., Isbell, D., May, L. D. N., Gutierrez Arvizu, M. N., & Jamieson, J. (2017). Equating in small-scale language testing programs., *Language Testing, 34*(1), 127-144. https://doi.org/10.1177/0265532215620825 (First Published December 23, 2015)

Equating in small-scale language testing programs

Geoffrey T. LaFlair

Daniel Isbell

L.D. Nicolas May

Maria Nelly Gutierrez Arvizu

Joan Jamieson

Abstract

Language programs need multiple test forms for secure administrations and effective placement decisions, but can they have confidence that scores on alternate test forms have the same meaning? In large-scale testing programs, various equating methods are available to ensure the comparability of forms. The choice of equating method is informed by estimates of quality, namely the method that has the least error as defined by random error, systematic error, and total error. This study compared seven different equating methods to no equating—mean, linear Levine, linear Tucker, chained equipercentile, circle-arc, nominal weights mean, and synthetic. The equating methods were evaluated based on the amount of error they introduced and their practical effects on placement decisions. A non-equivalent groups anchor test (NEAT) design was used to compare two listening and reading test forms based on small samples (one with 173 test-takers; the other, 88) at a university's English for Academic Purposes (EAP) program. It was found that two types of error (systematic and total) could not be reliably computed due to lack of an adequate criterion; consequently, only random error was compared. Among the seven methods, the circle-arc method introduced the least amount of error as estimated by the standard error of equating (SEE). Classification decisions made using the seven methods differed from no equating; all methods indicated fewer students were ready for university placement. Although interpretations regarding the best equating method could not be made, circle-arc equating reduced the amount of random error in scores, has reportedly low bias in other studies, accounts for form and person differences, and is relatively easy to compute. It was chosen as the method to pilot in an operational setting.

Keywords: English for academic purposes, equating, listening, reading, placement, sample size

Equating in Small-scale Language Testing Programs

Testing programs typically administer alternate forms of a test at different times. These forms should be written to the same content and statistical specifications, and they should produce interchangeable scores. Indeed, testing standards oblige test-makers to provide evidence regarding the comparability of alternate forms and test scores (AERA, APA, NCME, 2014, p. 104; ILTA, 2007, p. 4). Such evidence is important because even though different forms of a test are developed using the same content specifications and are intended to measure the same abilities, the forms may vary in difficulty. Additionally, for each administration, the test-takers may vary in ability. Large scale testing programs address this issue through equating. Equating is a statistical procedure that adjusts for the difficulty between alternate forms and accounts for the ability levels of test-takers, allowing a score on either form to be used interchangeably. Equating, then, allows different forms of a test to be used with the confidence that the scores have the same meaning (Dorans, Moses, & Eignor, 2010; Kolen & Brennan, 2014; Livingston, 2004).

Should small-scale testing programs follow the lead of large-scale programs and use equating procedures to ensure that scores on alternate test forms are interchangeable? In the US, for example, many intensive English language programs try to follow best practices by developing their own in-house tests to place students into the program's specific levels of instruction and even into the university (Commission on English Language Program Accreditation, 2015, Student Achievement Standard 1). These small language programs often create different forms of their placement test to guard against potential exam coaching and enhance test security. Do scores on these different forms have the same meaning?

While language testers in small-scale programs do endeavor to create multiple forms built to the same set of specifications, several factors cast doubt on the true equivalence of forms. First, and perhaps foremost, different test forms feature somewhat different content. Topical content of language tests is thought to interact with test-taker background knowledge, affecting scores (Clapham, 1996). This is generally considered a problem in language testing, as topical knowledge is not what we wish to make inferences about. This is not the case for other tests (e.g., medical certifications), where topical knowledge is central to the construct.

Second, language tests often feature testlet-based design, where forms are composed of input texts and a number of corresponding questions. When generating a new form, testlets are selected from a bank to meet general test specifications; however this may result in forms with different numbers of total items, making score comparisons more difficult. This too may distinguish language tests from some other tests.

Finally, small-scale testing programs often face practical limitations on quality control measures. New testlets are piloted, but may only yield item statistics based on fewer than 100 test takers. After a round of revision, practical demands may require the testlet to be put into operational use, with limited information about empirical item/testlet difficulties.

Although the issue of equating in language programs was relevant for our context, it has not been extensively reported on in our professional literature. A search of *Language Testing* and *Language Assessment Quarterly* yielded zero results pertaining to equating placement test scores in small-scale language programs. This is unfortunate, as small-scale, high-stakes testing contexts could potentially reap great benefits by ensuring consistent decisions across multiple test forms.

Perhaps one reason for this absence is that traditional equating methods require relatively large samples, and so incorporating equating in small language programs may have been thought to create more problems than it would solve—small sample size has been associated with large estimates of random error. Within the past 10 years, several promising equating methods focusing on small-sample equating have been introduced in experimental studies through the use of resampling. That is, small numbers of test-takers were extracted from very large data sets, or large datasets were simulated based on known item parameters (Babcock, Albano, & Raymond, 2012; Kim, von Davier, & Haberman, 2008; Livingston, & Kim, 2009, 2010, 2011; Sunnassee, 2011). It remains to be seen whether these new equating methods can be practically applied to small data sets from language programs. The present study was conducted in a real-world, small-scale testing program to explore the viability of several different equating methods.

Overview of Equating

In order to understand better the potential for small-sample equating with only small data sets, this brief overview of equating includes an explanation of design options and choices among methods. Equating properties, quality, sample size, and group ability vs. form difficulty are also considered. All of these issues are explained in detail in Kolen and Brennan's (2014) authoritative book, *Test Equating, Scaling, and Linking*.

Equating designs. One of three different data collection designs is used in equating; each design has unique advantages. First, Random Groups involves two randomly equivalent groups of test-takers. Each group takes a different form of the test. This design can be implemented in regular test administrations, and many forms can be equated at once (Kolen & Brennan, 2014). Second, Single Group with Counterbalancing uses one group of test-takers who are divided into two subgroups (Kolen, 2007). Two forms of a test are administered in different orders to each

subgroup. As with the Random Groups design, this design allows for differences in forms to be attributed to form difficulty; although it requires fewer test-takers than the Random Group design, it requires a special administration. Neither of these two data collection designs are practical for small language programs, so the next design is particularly relevant. Common-item Nonequivalent Groups, the third data collection design, also requires fewer test-takers that the Random Group design. It uses two groups of test-takers at two different times. This design is used when only one form of the test can be administered on a certain testing date. The second form, given on another date, contains a set of common items that were also administered in the first form. These common items are referred to as anchors, leading to the designation, the NEAT design (Non-Equivalent groups Anchor Test; von Davier & Kong, 2005). It is important that the anchor items represent the items in the full test in content and statistical distribution. Two variations of this design depend on whether the anchors are internal and counted in the score, or external and not counted in the score. Unlike the first two designs, the test-takers of the two forms cannot be assumed to be equivalent in ability, so performance on the anchor items is used in the equating formula to adjust for group ability differences. An obvious advantage of this design is that only one form is administered at a time (Kolen, 2007).

Equating methods. Equating methods can be classified into two main types—traditional, commonly used for large-scale samples, and alternative, recently used for small-scale samples. Three traditional equating methods are mean, linear, and equipercentile. Item Response Theory (IRT) equating can also be considered among the traditional methods; while powerful, IRT requires sample sizes and technical expertise beyond most small-scale language programs, and so was not considered further. More recently, alternative methods have been proposed specifically for small samples, namely the circle-arc (e.g., Livingston & Kim, 2009), nominal

weights mean (e.g., Babcock et al., 2012), and synthetic methods (e.g., Kim et al., 2008). All of these equating methods have been compared to no equating, referred to as *identity* equating as briefly described below.

Identity equating means that the scores on the two forms are assumed to have the same meaning about test-takers' abilities. Identity equating assumes no differences between forms or groups, and so a score on the new form (after any necessary scaling) is equivalent to the same score on the reference form. This also means that scores on forms with differences in difficulty, small or large, are considered equivalent. The identity method is the simplest, strongest (in terms of assumptions), and in small-scale language testing programs, likely the most commonly used equating method.

Mean equating uses the difference between the mean of the old form (i.e., the reference form) and the mean of the new form; this difference is a constant which is added to the scores on the new form. It assumes that the variance of the two forms are equal. In the NEAT design, however, variance and covariance are taken into account by a chaining method (Babcock et al., 2012). Mean equating has been recommended for use when sample sizes are small, when accuracy is most important near the mean, and when the ability levels of the two groups of test-takers are not too different (Kolen & Brennan, 2014).

Linear equating uses both the mean and standard deviation of the raw scores on both forms to calculate a linear function. Both mean equating and linear equating are characterized by a straight line, but whereas mean equating uses a constant to adjust all scores, linear equating allows scores to differ in varying amounts along the scale (Kolen & Brennan, 2014; Skaggs, 2005). Two widely used techniques for linear equating are the Tucker and Levine procedures (see Kolen & Brennan, 2014, Chapter 4; Sunnassee, 2011, pp. 28-31).

Equipercentile equating is characterized in terms of the shape of the distributions—a curve, rather than a straight line. As Livingston (2004, p. 17) explained, equipercentile equating transforms “each score on the new form to the score on the reference form that has the same percentile in that group.” The forms can vary in different amounts in different places along the distribution. When the shape of distributions appear irregular, a technique known as smoothing is used to regularize the shape and reduce sampling error (Kolen & Brennan, 2014). When the NEAT design is used, the chained equipercentile method involves equating the new form scores to the scores on the common items, and then the scores on the common items are equated to the reference form (Albano, 2014a; Kolen & Brennan, 2014; Livingston, 2004). This method has two advantages: the two groups of test-takers do not need to be very similar in ability for this method to be accurate (Kolen & Brennan, 2014, Livingston, Dorans, & Wright, 1990); and, it is less intensive to compute than other NEAT equipercentile designs. One disadvantage of chained equipercentile equating is that tests of very different lengths are not interchangeable.

Circle-arc equating was developed especially for small samples (Livingston & Kim, 2009). The method assumes a curvilinear relationship and applies a geometric transformation (namely, the arc of a circle) to a linear relationship between the score ranges of the reference form and the new form. Circle-arc equating utilizes three points (low, mid, and high) to define the relationship between the reference form and the new form (Babcock et al., 2012; Livingston and Kim, 2009). Advantages of the circle-arc method include suitability for a small sample size and ease of computation. It requires only the total number of items on each form, the means for each form, the means of the anchor set for each group of examinees, and the standard deviations of the reference form and the anchor set for the first set of examinees.

Nominal weights mean equating can be considered a simplification of Tucker linear equating (Babcock et al., 2012). Instead of employing variance and covariance, as is commonly done in other equating methods, nominal weights mean equating uses a ratio of the number of total form items to the number of anchor items to effectively "scale up the difference between the two groups' performance" (Babcock et al., 2012, p. 614) on anchor items. It is argued that this is beneficial when samples are small, as variance estimates may be unreliable and could introduce error in equating relationships.

Synthetic equating presents a best-of-both-worlds compromise between doing no equating vs. a given equating method, since no equating (i.e., identity) introduces no random error whereas equating methods are thought to account for form and/or test-taker differences (Kim, et al., 2008). As such, a predetermined weight is assigned to a given equating method, and its converse is applied to the identity method. This results in a partial transformation of scores based on the chosen equating method. Kim, et al. (2008) evaluated synthetic equating using chained linear equating, and Babcock et al. (2012) used linear Tucker, though any equating method could conceivably be entered into a synthetic equating function.

Issues affecting equating. Apart from the design and method used, equating requires that certain properties be met, that error be considered as a mark of quality, that cautions be acknowledged regarding sample size, and that differences in group ability and form difficulty be properly examined.

Properties. Five properties are considered when two test forms are equated (Holland & Dorans, 2006; Kolen & Brennan, 2014). The first two properties are considered as preconditions for equating. The same specifications property requires that the two forms be developed to the same test specifications and statistical specifications; in other words, the two forms are

measuring the same construct and have equal reliability. The symmetry property requires that the function used to transform a score on Form A to Form B must be the inverse of the function used to transform a score on Form B to Form A; this property precludes the use of regression techniques for equating. The next three properties are desirable, but are sometimes more difficult to examine. The equity property holds that it should not matter to a test-taker whether he takes Form A or Form B after equating. The observed score equating property addresses which statistical characteristics should be the same on two equated forms; the specific characteristics vary depending on the equating method. Finally, the group invariance property holds that the equating function should be the same across subgroups of test-takers from the same population. The degree to which these properties hold affect the confidence one can have in the equating results.

Error. The quality of equating methods is compared using error terms. The more error, the less confidence one can have in the equating method. Error is conceptualized as either random error (standard error of equating, SEE), systematic error (bias), and/or total error (root mean squared error, RMSE) (Kolen & Brennan, 2014; Motika, 2003; Sunnassee, 2011). The standard error of equating is inversely related to the sample of test-takers for each form; it decreases as sample size increases (Kolen & Brennan, 2014; Livingston, 2004). Bias—or systematic error—associated with the equating method is the difference between the estimated equated relationship and a criterion (or *true*) equating relationship. Total error—or Root Mean Squared Error, RMSE—is SEE and bias combined. Previous studies have generally been conducted within a theoretical test model which includes the notion of a true score or latent trait. They have relied on large scale data, using either the synthetic mean from both groups combined or the equipercentile equated scores to establish a criterion. Due to small sample sizes, this study

focused on observed score equating which is not dependent on a particular model. Without the notion of a true score or latent ability, there is no known way to establish an adequate criterion for bias or RMSE in these small-scale situations (M. Kolen, personal communication, December, 2014).

Sample size. As interest in small sample equating has grown, researchers have claimed that equating can be used with fewer test-takers. When discussing small-samples, sizes of 25-400 test-takers are the most common operationalization (see Heh, 2007, p. 25-31 and Livingston & Kim, 2009, p. 331 for reviews of small sample sizes in equating). This was not the case when considering traditional methods; Kolen and Brennan (2014) as well as Skaggs (2005) and Heh (2007) recommended samples of 400 for mean, linear, and Rasch methods and 1500 for equipercentile and 3-parameter IRT methods.

Focusing on NEAT designs, researchers have provided evidence that equating may be feasible for sample sizes as low as 50. Babcock, et al. (2012) found that sample sizes of 50 were viable for mean, linear, and nominal weights mean equating. Comparing identity, chained linear, and a synthetic equating of the two in a NEAT design, Kim, et al. (2008) found synthetic equating to outperform identity when sample sizes were 50 or greater. Livingston and Kim (2009) found that the circle-arc method outperformed traditional and identity methods in a NEAT design with a reference form sample of 75 and new form sample of 25. Research on the utility of smoothed equipercentile methods using small samples has shown mixed results. Livingston (1993) and Babcock, et al. (2012), using NEAT designs, found that presmoothing up to three moments in equipercentile equating can substantially reduce the amount of systematic error in samples as small as 50 cases when the two test forms substantially differ in difficulty.

However, in these studies and others using the random groups design with small samples (e.g., Livingston & Kim, 2010), equipercentile equating has not been found to perform as well as alternative methods. Using a random-groups design with small samples, Skaggs (2005) reported that mean equating produced the least error with sample sizes that were 50 and above when passing scores were below the mean.

Difficulty and ability. Differences in form difficulty and group ability can result in equating relationships that are inconsistent across methods (Babcock, et al.; 2012; Powers & Kolen, 2014; Sunnassee, 2011). For very small samples, identity equating has been recommended if other methods of equating introduce more error (Kolen & Brennan, 2014; Skaggs, 2005), especially if the forms are nearly equal in difficulty; however, identity becomes more inaccurate as the difficulty between forms increases (Babcock, et al., 2012). Generally, more disparate group abilities can result in inaccurate and varied equating results and a violation of assumptions (Powers, 2010; Sunnassee, 2011). Violations of assumptions, such as equal group ability in mean equating, increases bias. Babcock et al. (2012) conducted three sets of simulation studies to investigate the performance of seven different equating methods when sample size and form difficulty differed. The results showed that several small-scale equating methods performed well in situations where there were differences in group ability and form difficulty.

The Study

Because of our need for interchangeable scores on alternate forms of the placement test in our English for Academic Purposes (EAP) program, we investigated the difference between no equating and equating. Two main research questions were addressed: Which methods would

introduce acceptable levels of standard error into the equating process? What effects did equated scores have on placement decisions?

Method

A non-equivalent anchor test (NEAT) design was used to compare seven different equating methods to identity equating. In this section, the context, the test data, and the procedures used are described.

Context

An EAP program at a university in the southwestern United States administered a battery of tests to international students (non-native speakers of English) in order to determine their proficiency level in English upon arrival. The placement test battery had four parts: listening, reading, speaking, and writing. Listening and reading skills were tested using objective, dichotomously-scored multiple choice items. Speaking and writing skills were tested using performance-based tasks scored holistically by two trained raters. The total scaled scores were used to place students into one of five levels within the EAP program ranging from beginner to advanced, or to exempt students from intensive English study and allow matriculation to the university.

Test Data

Data from two placement test batteries administered in Fall 2011 and Fall 2012 were used. The listening and reading tests were equated. The speaking and writing tests were not equated but are described below and considered later in the procedures when examining placement decisions.

The speaking test on each form consisted of three constructed-response tasks. The first was telling a story based on pictures, the second was expressing an opinion, and the third was

summarizing a listening passage and giving an opinion. Task order, time constraints, and rubrics were consistent; topic and input differed slightly. Two raters scored each task using a 4-point holistic rubric; raters' scores were summed, scores on the three tasks were totaled, and then scores were scaled to 30. The mean of the 2011 form was 10.91 (*SD* = 4.96). The mean of the 2012 form was 12.89 (*SD* = 4.73).

The writing test on each form also consisted of constructed-response tasks—two (2011) or three (2012). The first task was writing a summary of a chart, the second was writing an opinion essay, and the third (for the 2012 form) was writing an email to a professor. The first two tasks were presented in the same order, given the same amounts of time for completion, and scored using the same rubrics for both administrations (barring topic-specific rubric notes), though topics and input differed slightly. On the 2012 form, the email task was presented last. Two raters scored each task using a 5-point holistic rubric; raters' scores were summed, scores on the two or three tasks were totaled, and then scaled to 30. The 2011 form had a mean of 12.90 (*SD* = 8.00), and the 2012 form had a mean of 16.93 (*SD* = 5.07).

The different listening and reading test forms were designed according to the same specifications. The items were grouped by passages, on a variety of academic topics. They were designed to measure examinees' ability to understand vocabulary, main ideas, detailed information, text organization, and inferences. The Fall 2011 test battery had 30 listening items and 35 reading items; these served as the *reference* forms. The *new* forms, administered in Fall 2012, had 35 listening items and 35 reading items.

Internal anchor sets of items were used in the listening and reading tests for both administrations. The listening anchor set comprised 9 items from two listening testlets—a

conversation between two students and a lecture about economics. The reading anchor set consisted of 11 items from one testlet on the topic of bioluminescence.

The Fall 2011 test was administered to 173 students. The Fall 2012 test was administered to 88 students. In both administrations, the majority of test-takers were from Middle-Eastern and Asian countries and were considered to represent the target population of the EAP program.

Descriptive and item statistics as well as reliability estimates were computed for Fall 2011 (reference forms) and Fall 2012 (new forms), as shown in Table 1.

Table 1

Test Statistics for Reference and New Listening and Reading Forms

	Listening		Reading	
	Fall 2011 Reference	Fall 2012 New	Fall 2011 Reference	Fall 2012 New
Number of examinees (*N*)	173	88	173	88
Number of items (*K*)*	30	35	35	35
Test Statistics				
Mean (*M*)	14.01	20.84	14.64	21.14
Standard Deviation (*SD*)	4.77	6.25	5.68	5.87
Cronbach's alpha (α)	0.73	0.83	0.79	0.83
Item difficulty (*p*)	0.47	0.60	0.42	0.60
Anchor Set Statistics				
Number of items (*K*)	9	9	11	11
Mean (*M*)	4.41	4.91	4.06	4.43
Standard Deviation (*SD*)	1.80	1.96	1.92	2.37
Item difficulty (*p*)	0.49	0.55	0.37	0.40
Correlation with subsection (*r)*	0.76	0.79	0.76	0.82

**Note.* The listening forms had different numbers of items.

Because the listening forms had different numbers of items, the means and standard deviations are not easily compared. Looking instead at the average item difficulty statistics, the reference form (.47) appeared to be more difficult than the new form (.60). Cronbach's alpha for the listening reference and new forms were .73 to .83 respectively, indicating that these forms had fairly typical internal consistency for tests not created by large-scale testing programs (Subkoviak, 1988). Performance on the listening anchor set also suggested a difference in group ability as students scored somewhat lower on the reference form than the new form and the standardized mean difference was 0.37 (greater than .25; Powers & Kolen, 2014).

Both reading forms had 35 items. Cronbach's alpha for the reading reference and new forms were .79 and .83. The means and standard deviations, and the average item difficulties suggested that the reading reference form was more difficult than the new form. Performance on the reading anchor set indicated smaller differences than listening in group ability with a standardized mean difference of 0.26. The correlation between the anchor sets and the overall results per section ranged from .76 to .82, indicating that the anchor sets were representative of the total tests.

In sum, it appeared that the listening reference form was more difficult than the new form and the students who took that form had less ability than the students who took the new listening form. It appeared that the differences in reading scores were mostly due to form difficulty.

Procedures

In this section, details about the equating methods used are briefly described, followed by an explanation of the estimation of standard error of equating. The steps taken to compare placement decisions are also described.

Equating methods. The new form was equated to the reference form for both listening and reading. Seven different equating methods were examined in addition to identity (no equating): (a) mean, (b) linear Levine, (c) linear Tucker, (d) pre-smoothed (to three moments) chained equipercentile, (e) circle-arc, (f) nominal weights mean (i.e., 25%), and (g) synthetic equating (Babcock, et al., 2012; Kim et al., 2008; Kolen & Brennan, 2014; Livingston 1993; Livingston & Kim, 2009). Mean and equipercentile equating used the chained equating method. For equipercentile equating, the score distributions for the new form, reference form and anchor sets were presmoothed to three moments to adjust for possible score sampling error that could be introduced by variation of scores in the distributions (Kolen & Brennan, 2014); this decision was based on an examination of the score distributions. For circle-arc equating, the low point was set at the chance score. Nominal weights mean equating considered the ratios of the number of anchor items to total items and mean scores. For synthetic equating, identity equating was paired with chained linear equating and the weight was set at 0.50, following Kim et al. (2008).

Equating error. The standard error of equating (SEE), the random error introduced by an equating method, was computed conditionally (i.e., at each score point) and as an overall average. As discussed previously, although equating error is usually examined in terms of random error (standard error of equating; SEE), systematic error (bias), and a composite of the two (root mean squared error; RMSE), bias and RMSE could not be calculated due to the lack of a population equating relationship necessary to establish a criterion.

SEE values were estimated using the results of bootstrapped equated relationships. Bootstrapping is a resampling procedure that creates “new” datasets by randomly resampling observations with replacement from the original data (Albano, 2014a; Davison & Hinkley, 1997). This differs from the resampling techniques used in other studies which have a large

population (typically 6,000 to 10,000 test-takers) from which to draw their samples. Those studies used simple random sampling without replacement; thus, within each of their replicated datasets, all of their observations were unique (c.f., Kim & Livingston, 2010; Livingston & Kim, 2009, 2010). Despite these differences in resampling, the calculation of the error estimates is the same.

The resampling analyses for each equating method consisted of 1,000 replicated datasets. The replicated datasets were the same size as the original datasets—new form (N = 88) and reference form (N = 173). With each replicated dataset the equated relationship between the new and reference form and its errors were estimated. Each of these equating methods and the equating errors associated with them were computed using the *equate* package version 2.0-3 in R (Albano, 2014a,b; R Core Team, 2014).

In the formulas for error below, the notation for each method can be interpreted as follows:

i represents the new form score points (0-35 in listening and reading).

j represents the bootstrap replication from 1-1,000 and its accompanying equating procedure.

K represents the number of items on the new form.

x_{ij} represents the equated score at level i for the new form on the jth bootstrap replication.

$\bar{x}_i$ represents the mean of x_{ij} over 1,000 bootstrap replications.

The random error at each score point in the series of possible scores (conditional SEE) was estimated by calculating the difference between the mean equated value for each score point

and the estimated value in the *j*th replication. Then, at each score point in the series of possible scores, the difference was squared and all squared differences were summed. Finally, the square root of the summation was calculated (1). This number represented the random error at the score points of the equating relationship.

$$SEE_i = \sqrt{\frac{1}{1000}\sum_{j=1}^{1000}\left(x_{ij} - \bar{x}_i\right)^2} \quad (1)$$

The mean SEE at the raw score level was calculated by summing these errors at each score point and averaging them over the number of items on the new form (2). This gave an estimate of the mean random error across all score points on the raw score scale.

$$\overline{SEE} = \frac{1}{K}\sum_{i=1}^{K}\sqrt{\frac{1}{1000}\sum_{j=1}^{1000}\left(x_{ij} - \bar{x}_i\right)^2} \quad (2)$$

The mean SEE was interpreted as the average range of variability in terms of raw score points on the reference form of the equating method over replications.

Placement decisions. In order to examine the effects of equating on placement decisions, the total scores on the placement battery were examined for each of the equating methods. Each test (reading, listening, writing, and speaking) was scaled to 30 points via a constant multiplier (30 divided by the maximum possible score for each test); once summed, a total score resulted, ranging in value from 0 to 120. This total score was used to determine the placement of each test-taker. Admission to the university was based on a cut score of 70. Scores below 70 required a student to take preparatory English courses. This cut score represents the highest stakes for test takers and test users, and as such is the primary focus in consideration of placement decisions. Placement into 5 different levels of the EAP program was based on the following cut scores: 0-15 for Level 1; 16-31 for Level 2; 32-44 for Level 3; 45-56 for Level 4; and 57-69 for Level 5. These cut scores may be considered as having lower stakes, as placements within (and only

within) the program could be appealed after a week of class and students had the opportunity to re-test for EAP promotion/university matriculation after a semester of study. Both number and percentage of students at each level were computed. The percentages of the Fall 2012 form for each equating method were compared to those of the Fall 2011 (reference form) results as a consistency check (Kolen & Brennan, 2014).

Results

Evaluation of the equating methods showed that circle-arc introduced the least amount of random error (SEE) through most of the score range for each test and overall. Identity equating introduced no random error, as would be expected since no changes were made to the scores. Classification decisions made using the seven methods differed from no equating; all methods indicated fewer students were ready for university placement.

Comparison of SEE

Figure 1 shows the conditional SEE (i.e., random error) for listening (top) and reading (bottom) at each score point for the equating methods. The raw score scale for the new form is on the x-axis, and the estimated error is on the y-axis. The error values on the y-axis can be interpreted on the same scale as score points.

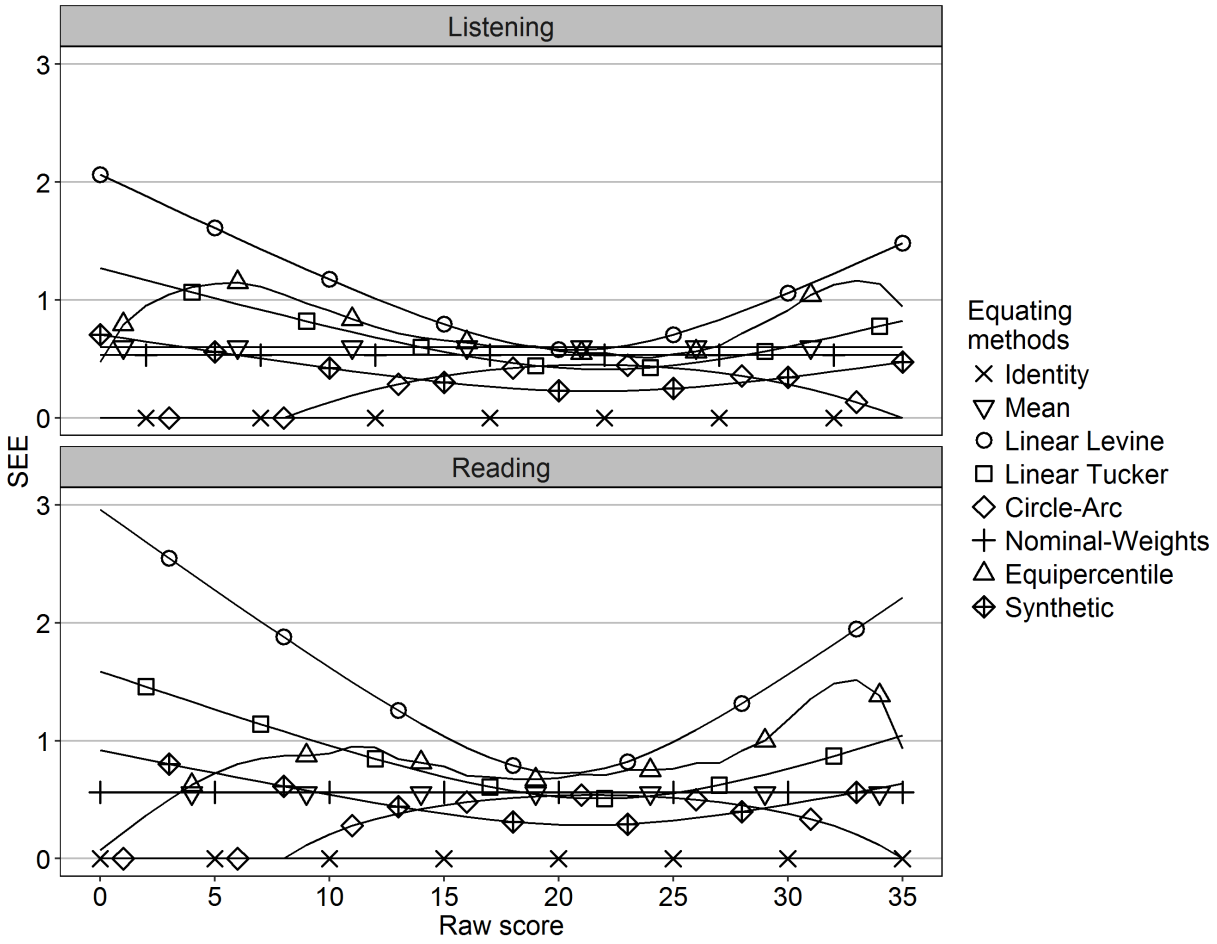

Figure 1. Standard error (SEE) values of the equating methods for the listening and reading tests.

Ideally, the error for each score point would be at or near zero. Looking closely at the conditional SEE values for each of the methods, identity had no standard error at any score point. This was expected because this method did not make any adjustments to scores. For example, a score of 5 on the new form would always be equivalent to a score of 5 on the reference form. Next, mean and nominal weights mean equating had the same flat line as identity because the equated score adjustment was the same for each raw scale point; however, they both had error over replications. The two linear methods (Levine and Tucker) showed a u-shaped error pattern with more error at the low and high ends of the scoring scales than in the middle. Scores below

chance (the low point) of circle-arc did not vary, as they were defined by a fixed relationship. The circle-arc method had low random error throughout the range of scores, tapering off to zero near the low and high points, highlighting the method's constraints on the relationship between forms. Finally, both synthetic and equipercentile equating showed larger amounts of conditional SEE at the high and low ends of the scoring scale. Synthetic equating resulted in relatively larger amounts of error at score points below 15 and above 25 for both the reading and the listening tests. Also, most of the random error was at the ends of the score scale for the equipercentile method. Notably, only linear Tucker, Levine, and equipercentile had conditional error greater than 1 at any score point on either test.

Overall (excluding identity), the various methods demonstrated similar magnitudes of conditional SEE in the middle of the score range, but differed somewhat at the ends. Also, the SEE for each method showed similar patterns across the listening and reading tests.

The results of these analyses at the individual score points can be summarized by estimating the mean SEE across all items on the new form. The SEE averages for each test using the equating methods are shown in Table 2. The circle-arc method had the lowest average SEE, apart from identity. Linear Levine had the greatest amount of SEE followed by linear Tucker and equipercentile. Mean and nominal weights mean equating had similar mean SEE estimates on both tests. Synthetic equating's SEE was close to the circle-arc methods on the listening test, but was more similar to mean and nominal weights mean on the reading test.

Table 2

Mean SEE for Equated Listening and Reading Tests

Method	Listening	Reading
Identity	0.00	0.00
Mean	0.58	0.56
Linear Levine	1.18	1.56
Linear Tucker	0.70	0.87
Circle-Arc	0.23	0.29
Nominal Weights Mean	0.53	0.55
Equipercentile	0.82	0.83
Synthetic	0.40	0.51

Effects on Placement Results

To compare the practical effects and examine the consistency of the different equating methods, bar charts were generated to display the percentage of students placed at each level in the EAP for the reference form (Fall 2011; at the top of Figure 2) and for the new form (Fall 2012) based on identity and the 7 equating methods.

First, if we consider placement into the university vs. the intensive language program one can see that on the reference form 31.79% of the test-takers scored in the band at the far-right (representing admission to the university).

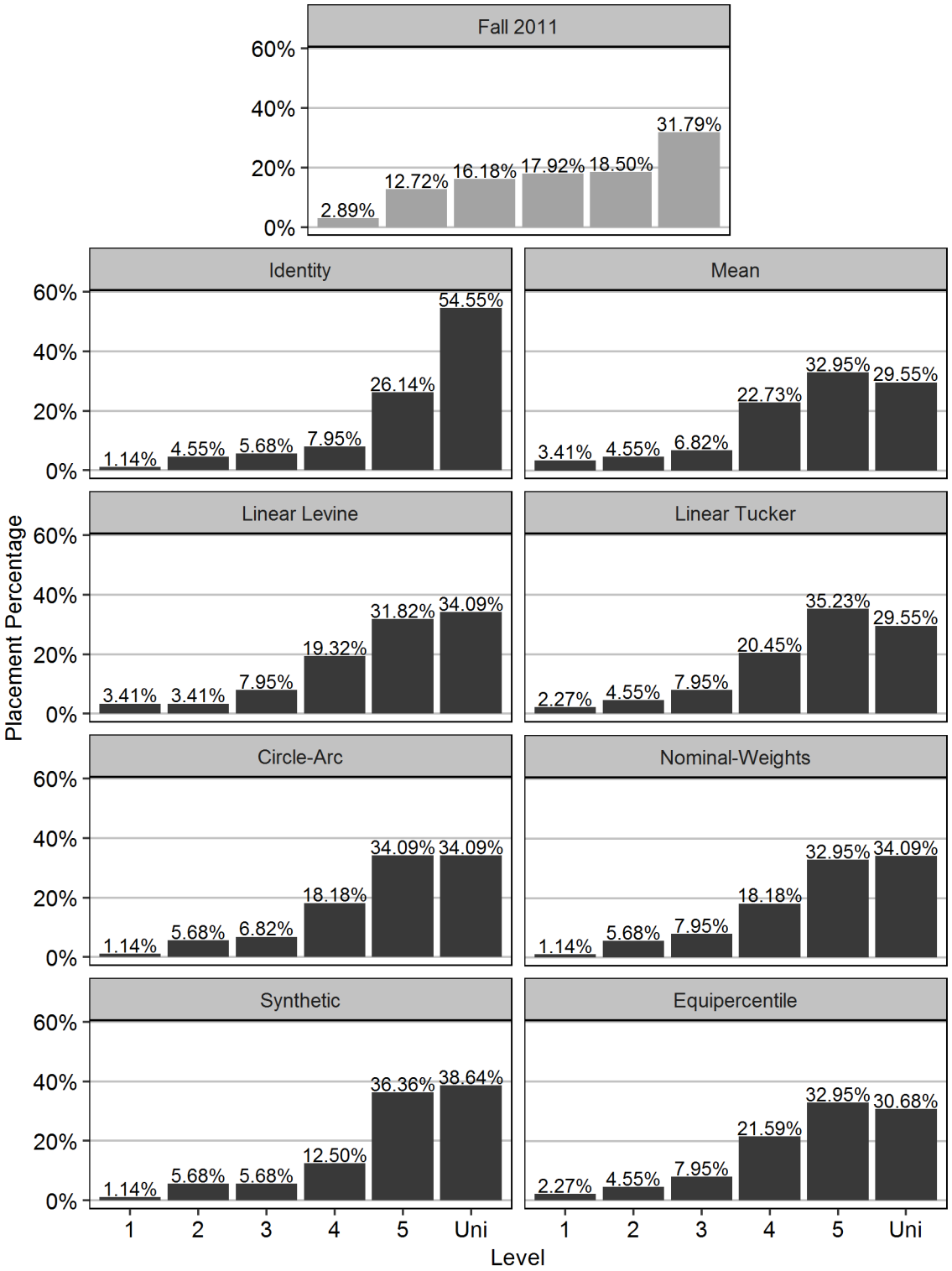

Figure 2. Examinee placement based on equating method and compared to reference form.

The identity method placed more than half of test-takers (55%) into the university. All of the other equating methods resulted in about 20% fewer placements at that level. Synthetic equating placed the second greatest proportion of students in the university (38.64%). Linear Levine, circle-arc and nominal weights mean displayed similar decision profiles, placing 34.09% at the university level. Equipercentile placed 30.68%. Mean and linear Tucker equating were the most conservative, placing only 29.55% of students in the university. If we consider placement at the two highest levels (university and Level 5) a similar pattern emerges. The greatest amount of students were placed here by identity (80.69%), followed by synthetic equating (75%). The remaining equating methods resulted in about 10-15% fewer placements at these levels: circle-arc (68.18%), nominal weights (67.04%), linear Levine (65.91%), linear Tucker, (64.87%), equipercentile (63.63%), and mean (62.5%). The equating methods showed little difference in placement decisions at Levels 1, 2, and 3, ranging from about 11% to 15%.

Conclusion

Analysis of the listening and reading forms of the test battery revealed differences in difficulty and test-taker ability. These differences suggested that test equating may be beneficial for consistent placement decisions. Analysis of seven equating methods revealed little random error, with the circle-arc method introducing the least amount of SEE. Moreover, because they accounted for group differences, circle-arc, nominal weights mean, and synthetic equating appeared to be most appropriate. Analyzing the placement decisions showed major contrasts between equating and no equating.

Comparing the equated placement distributions with the Fall 2011 (reference form) group showed that all equating methods, excluding identity, were consistent with the reference form in the sense that they placed a similar percentage of examinees into the university—the decision

point with the highest stakes. Given the apparently greater ability of the new form group, it seems reasonable that most of the equating methods (all but synthetic and identity) placed a higher proportion of examinees into the university and Level 5 than the reference group. Despite the apparent strength of the new form group, identity and synthetic equating seem to have disproportionately placed students into the higher levels, and thereby reduced placements in several of the language program's lower levels.

Before discussing implications of these findings, important limitations must be acknowledged. Primarily, a full evaluation of the quality of equating methods was not possible in this study. While the low SEE values are promising considering the small samples, the lack of a criterion equating relationship prevented rigorous analysis of equating bias. This makes it impossible to identify the best equating method, as it is quite possible for a method with low SEE to have high bias, and vice-versa. Considering group abilities, placement consistency, and potential violations of assumptions only allows for coarse, context-specific inferences on equating quality to be made. Without empirical bias values, the picture is incomplete.

In this study, equating methods were compared by their amounts of random error, or SEE. The circle-arc method had the least overall SEE; in fact, all equating methods had rather low overall SEE. To put the mean SEE values found in the present study into perspective, consider the standard error of measurement (SEM) associated with each of the new form (Fall 2012) tests: 2.57 for Listening and 2.42 for Reading. Even the largest SEE values, generated by the linear Levine method, were less than half of the magnitude on average of the SEM; however, the two linear Levine did exhibit error estimates as large as, or larger than, the SEM at the lower and upper ends of the score scale for the reading test.

This finding of relatively little random error is important, as issues with sample size have often been seen as a major obstacle to small-scale equating. The small SEE values obtained in the present study are congruent with results from recent research into circle-arc equating (Livingston & Kim, 2009; Livingston & Kim, 2010), nominal weights mean equating (Babcock et al., 2012), synthetic equating (Kim et al., 2008), and seminal research involving small-sample equipercentile equating (Livingston, 1993). Given the small SEE values found in the present study, one might speculate that bias and overall error values would be similarly small as indicated in results obtained in other studies. This speculation seems warranted if it is not possible to calculate bias and RMSE for small-scale observed score data due to lack of a readily available criterion. However, this leaves practitioners in language testing programs relying on results from simulation and large-scale resampling studies to characterize probable bias and RMSE in small-scale equating relationships.

When bias is considered in this way, the circle-arc method has consistently shown comparatively less systematic error (i.e., bias) than other methods (e.g., Babcock, et al. 2012; Livingston & Kim, 2010). Recently, general linear methods for identity, mean, and linear equating have been shown to reduce bias in equating (Albano, 2015). This promising equating method has been used to account for differences in test difficulty due to differing scale lengths, violations of assumptions, or differing numbers of test takers between forms (Albano, 2015).

When RMSE is considered, alternative and traditional equipercentile methods have been found to be robust to large differences in group ability (Babcock et al., 2012; Powers & Kolen, 2014), and alternative, small-scale methods have also been found to perform strongly when there are both differences in group ability and form difficulty (Babcock et al., 2012). It must be reemphasized, though, that this is only optimistic speculation.

The second issue, the practical significance of equating methods, was examined by how they affected placement decisions. If equating were not used, a larger number of students may have been admitted to the university—potential false positives, in terms of admission decisions; all seven equating methods reduced that number. In that sense, all equating methods examined may appear preferable to no equating. It remains a possibility, however, that some (or all) of the methods created false negatives. Without information on bias based on a "true" criterion, it is impossible to know which different equating method's placement results would be closest to the "truth." The findings reported for this case study illustrate the placement decisions that different equating methods would yield, but we cannot identify the best method.

Taking all of this information into account, we have decided to pilot the circle-arc equating method in an operational setting. Score reports need to be generated quickly, but freely available software for computing equating relationships (*R* and the package *equate*) makes equating practical for small-scale language testing programs. For all of these methods, some familiarity with *R* and the *equate* package is necessary to quickly conduct test equating. The circle-arc method is also a viable option for EAP program staff who do not have expertise in R because it only requires familiarity with conventional mathematical order of operations and the availability of spreadsheet software to carry out equating in a reasonable amount of time.

Although we have arrived at a practical conclusion for our specific small-scale language testing context, concern remains. Babcock et al. (2012) and Kolen and Brennan (2014) recommended that no equating (i.e., identity equating) be used if forms were nearly equivalent in difficulty. This begs the question: When deciding to rule out equating due to similar difficulty, how close is close enough? Efforts by Skaggs (2005), Heh (2007), and Sunnassee (2011) have made progress toward guidelines, but further research in this area leading to a consensus for

when and how to conduct small-scale equating would be a boon for the field of small-scale language testing. It could lead to more transparent, interchangeable scores for placement decisions.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, D.C.: Author.

Albano, A. D. (2015). A general linear method for equating with small samples. *Journal of Educational Measurement, 52*, 55-69.

Albano, A. D. (2014a). *equate: Statistical methods for test equating* [Computer software manual]. http://CRAN.R-project.org/package=equate

Albano, A. D. (2014b). *equate: An R package for observed-score linking and equating. Version 2.0-3.* Retrieved from http://cran.r-project.org/web/packages/equate/vignettes/equatevignette.pdf

Babcock, B., Albano, A., & Raymond, M. (2012). Nominal weights mean equating: A method for very small samples. *Educational and Psychological Measurement, 72,* 608-628.

Clapham, C. (1996). The development of IELTS: A study on the effect of background knowledge on reading comprehension. *Studies in Language Testing, 4*. Cambridge: Cambridge University Press.

Commission on English Language Program Accreditation. (2015). *CEA Standards for English Language Programs and Institutions.* Retrieved from http://www.cea-accredit.org/about-cea/standards

Davison, A., & Hinkley, D. (1997). *Bootstrap methods and their applications.* New York, NY: Cambridge University Press.

Dorans, N., Moses, T., & Eignor, D. (2010). *Principles and practices of test score equating.* ETS Research Report RR-10-29. Princeton, NJ: Educational Testing Service.

Heh, V. (2007). *Equating accuracy using small samples in the random groups design.* (Doctoral dissertation. Ohio University). Available from ProQuest Dissertations and Theses database. (UMI No. 3275287)

Holland, P., & Dorans, N. (2006). Linking and equating. In R. Brennan (Ed.), *Educational measurement* (4th ed.). (pp. 187-220). Westport, CT: American Council on Education and Preager Publishers.

International Language Testing Association. (2007). *Guidelines for practice.* Retrieved from http://www.iltaonline.com/images/pdfs/ilta_guidelines.pdf

Kim, S., & Livingston, S. (2010). Comparisons among small sample equating methods in a common-item design. *Journal of Educational Measurement*, *47*, 286-298.

Kim, S., von Davier, A., & Haberman, S. (2008). Small-sample equating using a synthetic linking function. *Journal of Educational Measurement*, *45*, 325-342.

Kolen, M. (2007). Data collection designs and linking procedures. In N. Dorans, M. Pommerich, & P. Holland, (Eds.) *Linking and aligning scores and scales* (pp. 31-55). New York, NY: Springer.

Kolen, M., & Brennan, R. (2014). *Test equating, scaling, and linking* (3rd ed.). New York, NY: Springer.

Livingston, S. (1993). Small sample equating with log-linear smoothing. *Journal of Educational Measurement*, *30*, 23-39.

Livingston, S. (2004). *Equating test scores (without IRT).* Princeton, NJ: Educational Testing Service.

Livingston, S., Dorans, N., & Wright, N. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education*, *3*, 73-95.

Livingston, S., & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement*, *46*, 330-343.

Livingston, S., & Kim, S. (2010). Random-groups equating with samples of 50 to 400 test-takers. *Journal of Educational Measurement*, *47*, 175-185.

Livingston, S., & Kim, S. (2011). New approaches to equating with small sample sizes. In A. A. von Davier (Ed.) *Statistical models for test equating, scoring, and linking* (pp. 109-122). New York, NY: Springer.

Motika, R. (2003). *Effects of anchor item content representations on the accuracy and precision of small sample linear test equating*. (Doctoral dissertation. University of South Florida). Available from ProQuest Dissertations and Theses database. (UMI No. 3116435)

Powers, S. (2010). *Impact of matched samples equating methods on equating accuracy and the adequacy of equating assumptions.* (Doctoral dissertation. University of Iowa). Available from ProQuest Dissertations and Theses database. (UMI No. 3439324)

Powers, S., & Kolen, M. J. (2014). Evaluation accuracy and assumptions for groups that differ in performance. *Journal of Educational Measurement*, *51*, 39-56.

R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org/

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

Subkoviak, M. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement*, *25*, 47-55.

Sunnassee, D. (2011). *Conditions affecting the accuracy of classical equating methods for small samples under the NEAT design: A simulation study.* (Doctoral dissertation. University of

North Carolina at Greensboro). Available from ProQuest Dissertations and Theses database. (UMI No. 3473486)

von Davier, A., & Kong, N. (2005). A unified approach to linear equating for the nonequivalent groups design. *Journal of Educational and Behavioral Statistics*, *30*, 313-342.