Equating at a Small Language Program

Geoffrey T. LaFlair, Daniel Isbell, Maria Nelly Gutierrez Arvizu, L. D. Nicolas May,

& Joan Jamieson

Northern Arizona University

Abstract

Because small-scale intensive language programs routinely administer tests to measure students' language proficiency, interchangeable forms are needed for secure and effective placement decisions. Equating is a statistical procedure that allows different forms of a test to be used with the confidence that the scores have the same meaning. Different equating methods were evaluated through two research questions: Does equating introduce more error than it accounts for? What effects do equated scores have on placement decisions? A non-equivalent groups anchor test (NEAT) design was used to compare two listening and reading test forms (one administration of 173 test-takers, the other, 88). Seven different equating methods were evaluated—identity, mean, linear Tucker, linear Levine, equipercentile, and two variations of circle-arc. Based on the standard error of equating (SEE), equating bias, and root mean square error (RMSE), the most error was present with no equating (i.e., identity) and mean equating. The circle-arc zero method introduced the least amount of error in total and at each score point. Classification decisions for each method differed at the high end of the scale. Using equating reduced the amount of error in scores and reduced the potential for false-positive decisions. The study contributes to the literature on small-sample equating with its use of actual, small data sets.

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

Equating at a Small Language Program

Background

Testing programs typically administer parallel forms of a test at different times. These parallel forms are written to the same content and statistical specifications, and they should produce interchangeable scores. However, 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 and, for each administration, the test-takers may vary in ability. This variation among forms leads to questions of fairness, which need to be addressed (Standard 4.10, AERA, APA, NCME, 1999). Equating is a statistical procedure that adjusts for the difficulty between the forms and accounts for the ability levels of test-takers, allowing a score on either form to be used interchangeably (Kolen & Brennan, 2004).

Small-scale intensive language programs routinely administer a form of a test to determine the placement of students before instruction begins. Interchangeable test scores are an important consideration for consistent, effective, and confident placement decisions. Should small-scale language testing programs use equating procedures to ensure that scores from different test administrations are interchangeable?

Traditional equating methods require relatively large samples, and so have not been used in small language programs primarily because they have been thought to create more problems than they solve—small sample size has been associated with large error estimates. Recently, studies have indicated the potential of equating when using small samples (e.g., Babcock, Albano, & Raymond, 2012; Livingston & Kim, 2009; Sunnassee, 2011). These studies have derived small samples by resampling, that is, by extracting different numbers of test-takers from very large data sets (e.g., Kim, von Davier, & Haberman, 2008; Livingston, & Kim, 2010, 2011).

However, it remains to be seen whether these methods can be practically applied to small data sets.

Research Questions

Because of our need for interchangeable scores on different forms of the placement test in our intensive English program, we investigated which, if any, equating method might be best for our context. Two main research questions were addressed: Does the use of equating introduce more error than it accounts for? What effects do equated scores have on placement decisions?

Methods

Data from the placement test battery that were equated consisted of two tests (listening and reading) of two administrations (Fall 2011 and Fall 2012). The different test forms were designed according to the same specifications. The items were grouped by passages, forming testlets, 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 the Middle-East and Asian countries and were considered to represent the target population of the EAP program.

A non-equivalent anchor test (NEAT) design was used to compare seven different equating methods. The new form was equated to the reference form for both listening and reading. Seven different equating methods were used: (a) identity (i.e., no equating), (b) mean, (c) linear Tucker, (d) linear Levine, (e) pre-smoothed (to three moments) chained equipercentile (f) circle-arc with a low-point equal to the chance score (i.e., 25%), and (g) circle-arc with a low point of zero (Kolen & Brennan, 2004; Livingston 1993; Livingston & Kim, 2009). Three types of error were examined for each of the seven equating methods for both listening and reading tests. First, the standard error of equating (SEE) is the random error that is introduced by an equating method. Second bias—or systematic error—associated with the equating method is the difference between the estimated equated relationship and a criterion equating relationship. Third, total error—or Root Mean Squared Error, RMSE—is SEE and bias combined.

In order to determine the effects of equating on placement decisions, the scores on all four sections of the placement battery were scaled to 30 points; once summed, a composite score resulted ranging in value from 0 to 120. This composite score was used to determine the placement of each test taker. Admission to the university was based on a cut score of 70. Placement into 5 different levels of the EAP program was based on different cut scores.

Results

Evaluation of the seven equating methods in respect to error showed that the circle-arc zero method introduced the least amount of error at each score point and in total. Classification decisions for each of the methods differed most at the high end of the scale.

The averages of the three types of error for each test using the seven equating methods are shown in Table 1. Both circle-arc methods had the lowest SEE, apart from identity (no equating). Linear Levine had the greatest amount of SEE. Circle-arc zero had the least amount

of systematic error (bias), coming closest to theoretical true scores. Circle-arc chance had less bias than equipercentile on the listening test; this was reversed on the reading test. Identity and mean had the largest bias. Circle-arc zero had the lowest RMSE on both tests. Circle-arc chance had less RMSE than equipercentile on the listening test, followed—from smallest RMSE to largest—by the linear methods, and finally, identity and mean. On the reading test, equipercentile had less RMSE than circle-arc chance, followed by linear Tucker and mean, and finally linear Levine and identity.

Table 1

Mean SEE, Bias, and RMSE for Equated Listening and Reading Tests

	Listening			Reading		
Method	SEE	Bias	RMSE	SEE	Bias	RMSE
Identity	0.00	1.50	1.52	0.00	4.01	4.01
Mean	0.69	-2.38	2.65	0.76	-2.47	2.69
Linear Levine	1.06	-0.66	1.35	1.57	-2.29	3.76
Linear Tucker	0.70	-1.21	1.45	0.89	-2.06	2.41
Equipercentile	0.95	-0.55	1.17	0.94	-0.45	1.15
Circle-Arc Chance	0.23	0.25	0.65	0.29	1.22	1.56
Circle-Arc Zero	0.31	-0.22	0.40	0.37	0.28	0.50

To compare the practical effects of the different equating methods, Table 2 details the distribution of placement results. The results indicated that the equating methods could be placed into three groups. First, the identity method placed more than half of test takers (55%) into the highest level (i.e., university); all of the other equating methods resulted in about 20% fewer placements at that level. Second, mean, linear Tucker, and linear Levine methods were similar. These methods had the fewest university placements (27% to 32%); linear Tucker had the most Level 4 placements (23.9%).Third, both circle-arc methods and the equipercentile method resulted in slightly over one-third of test-takers being allowed to enter the university, about 18%

placed in level 4, and about 30% placed in Level 5. The equipercentile method and the two circle-arc methods showed substantial agreement with the criterion method (Landis and Koch, 1977).

Table 2

Student Placement based on Equating Method

Method		Level 1	Level 2	Level 3	Level 4	Level 5	University	Cohen's κ
Identity	n	1	4	5	8	22	48	0.44
	%	1.1	4.5	5.7	9.1	25.0	54.5	
Mean	n	2	4	8	16	30	28	0.60
	%	2.3	4.5	9.1	18.2	34.1	31.8	
Linear Levine	n	3	3	9	18	29	26	0.56
	%	3.4	3.4	10.2	20.5	33.0	29.5	
Linear Tucker	n	3	3	8	21	29	24	0.58
	%	3.4	3.4	9.1	23.9	33.0	27.3	
Equipercentile	n	1	5	8	16	25	33	0.66
	%	1.1	5.7	9.1	18.2	28.4	37.5	
Circle-Arc chance	n	1	5	7	15	29	31	0.67
	%	1.1	5.7	8.0	17.0	33.0	35.2	
Circle-Arc zero	n	1	5	8	16	27	31	0.66
	%	1.1	5.7	9.1	18.2	30.7	35.2	

Note. N = 88

Relevance to PIE

Moving forward, we plan to pilot the circle-arc zero 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 a viable option for EAP programs that do not have expertise in R because it only requires familiarity with mathematical order of operations and the availability of spreadsheet software to carry out equating in a reasonable amount of time.

References

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

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

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., & Brennan, R. (2004). *Test equating: Methods and practices.* New York, NY: Springer.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics,* 33, 159-174.

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

Livingston, S., & Kim, S. (2009). The circle-arc method for equating 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.

R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org/

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)