Linking With External Covariates: Examining Accuracy by Anchor Type, Test Length, Ability Difference, and Sample Size

Anthony D. Albano

University of Nebraska–Lincoln

Marie Wiberg

Department of Statistics, USBE, Umeå University, Sweden

Author Note

Corresponding Author: Anthony Albano, 114 Teachers College Hall, Lincoln, NE, 68588, 402-472-8911, albano@unl.edu .

This research was funded in part by the Swedish Research Council Grant 2014-578.

This manuscript has been accepted for publication in *Applied Psychological Measurement*, DOI: 10.1177/0146621618824855.

Abstract

Research has recently demonstrated the use of multiple anchor tests and external covariates to supplement or substitute for common anchor items when linking and equating with nonequivalent groups. This study examines the conditions under which external covariates improve linking and equating accuracy, with internal and external anchor tests of varying lengths and groups of differing abilities. Pseudo forms of a state science test were equated within a resampling study where sample size ranged from 1,000 to 10,000 examinees and anchor tests ranged in length from 8 to 20 items, with reading and math scores included as covariates. Frequency estimation linking with an anchor test and external covariate was found to produce the most accurate results under the majority of conditions studied. Practical applications of linking with anchor tests and covariates are discussed.

Keywords: equating, linking, simulation

Linking With External Covariates: Examining Accuracy by Anchor Type, Test Length, Ability Difference, and Sample Size

A variety of methods are available for equating scores from one form of a test to another, so that the forms can be used interchangeably. In each method, differences in form difficulty are used to determine the equivalence, or lack thereof, between test forms. An equating method is effective when it accurately estimates the impact of form difficulty differences on score distributions while controlling for any differences in the groups taking each form and minimizing random and systematic errors. On the other hand, a method is ineffective when error overwhelms an estimated form difficulty difference, or when differences between groups are misestimated (von Davier, 2011).

Differences between the groups taking two test forms, X and Y, can confound estimates of form difficulty differences and must be isolated using an appropriate equating design. In single-group (SG) and equivalent-groups (EG) designs, samples are taken directly from the target population T, and are either the same or are considered to be randomly equivalent. In these designs, there are no confounding group effects. When equating with samples from potentially nonequivalent populations, P and Q, differences between the groups must be controlled for statistically. The nonequivalent-groups with anchor test (NEAT) design includes a common anchor test V that is administered to both groups, and scores on V are used to control for the difference between them. Although equating with the SG and EG designs tends to be simpler, involving fewer estimates and assumptions, the NEAT design is most often used in practice.

The objective in equating with nonequivalent groups is to statistically control for group differences without substantially increasing standard error and bias. Research has shown that equating with nonequivalent groups is improved when an anchor test is designed as a miniature version of the total tests, similar to them in content and certain statistical characteristics (e.g., Cook & Paterson, 1987; Klein & Jarjoura, 1985; Liu, Sinharay, Holland, Curley, & Feigenbaum, 2011; Sinharay & Holland, 2007). Studies have

also shown how equating with nonequivalent groups can be made more accurate when the traditional anchor test is augmented by other relevant measures of ability, including measures external to the test itself (e.g., Bränberg & Wiberg, 2011; Kim, Livingston, & Lewis, 2011; Liou, Cheng, & Li, 2001; Livingston, Dorans, & Wright, 1990; Wiberg & Bränberg, 2015). In these augmented equating methods it is assumed that equating error can be reduced through the inclusion of additional variables.

Previous research on equating with multiple common variables has primarily explored improvements in equating accuracy under two conditions: with and without one or more anchor variables or external covariates. For example, Bränberg and Wiberg (2011) showed that including information from covariates correlated with test scores, such as gender and education level, can reduce linear equating error, relative to equating without such information. Liou et al. (2001) used multiple imputation methods to reduce group nonequivalence prior to equipercentile equating in an NEAT design; multiple imputation using scores on another test as a surrogate variable produced results similar to NEAT equating with an internal anchor. Moses, Deng, and Zhang (2011) showed how equating with two anchor tests, one containing only multiple-choice items and the second containing constructed-response items, led to reductions in standard error of equating differences within the kernel equating framework, in comparison to equating with only the multiple-choice anchor. Wiberg and Bränberg (2015) compared equating in an EG design with equating under a NEAT design, a nonequivalent groups with covariate(s) (NEC) design, and a combination of the two where both an anchor test and covariate(s) can be used; standard errors were found to be smallest when all available variables, anchor and covariates, were used in equating.

More recently, Wallin and Wiberg (2017) proposed incorporating covariates via propensity scores within the kernel equating framework. Using data from Wiberg and Bränberg (2015) with total test scores coming from the quantitative section of a large-scale standardized test, standard errors were found to be lowest when equating in a NEC design

with propensity scores based on course grades and verbal test scores, compared with equating without anchor items. Sansivieri and Wiberg (2017) included covariates within an item response theory framework. Results from observed-score equating based on real and simulated data confirmed that standard errors were reduced when covariates were utilized in addition to anchor tests, compared with equating under EG and NEAT designs without information from covariates.

Studies on linking and equating with multiple common variables are limited in two main ways. First, they have not fully assessed equating accuracy for equipercentile methods in terms of overall error. A comprehensive evaluation of accuracy requires access to true equating functions, which are not typically available (as noted by Wiberg & Bränberg, 2015), and previous simulation and resampling studies have focused on small sample equating methods (e.g., Bränberg & Wiberg, 2011; Kim et al., 2011). Second, previous studies have yet to explore the different conditions under which linking and equating with multiple common variables are most useful. This paper addresses these two issues. Resampling studies, based on population data with known true equating functions, are used to examine linking and equating accuracy under conditions of varying sample sizes, examinee ability differences, and anchor test lengths, both with and without additional external covariates. Results provide insight into the benefits of incorporating external information into the linking process.

The term *multivariate* is used in this paper to denote equating methods that can incorporate multiple common variables measured for all examinees in a nonequivalent groups design. These variables may consist of one or more anchor tests along with external covariates that may or may not be designed to assess the same content or constructs as the total test. When the common variables come from instruments designed according to the same framework and outline as the total test, the methods can be expected to produce equating functions that generalize to the target population and that are invariant across subpopulations. For example, this should be the case in Moses et al. (2011), where an

anchor test was supplemented by scores on constructed-response items internal to the total test. Otherwise, common variables are expected to produce linking functions with which population invariance may not necessarily hold, as differences in subgroups on external covariates may not reflect differences on the total test. For example, this may be the case in Wiberg and Bränberg (2015), where supplemental information did not come directly from the total test. The term *equating* is used here unless a distinction between linking and equating is necessary.

Two equating methods, Tucker linear and frequency estimation equipercentile, are available in multivariate form (see Angoff, 1984; Gulliksen, 1950). The multivariate methods are generalizations of existing methods for NEAT equating with a single common variable, typically an internal or external anchor test, and in the presence of a single anchor they function the same (for details, see Albano, 2016a). The methods differ in their complexity and statistical assumptions. Frequency estimation is expected to require large samples (e.g., 1,000 or more) to produce accurate results. Including two or more additional common variables will increase the sample size requirements of the method, especially when these variables contain numerous score categories that can lead to sparser distributions (Wiberg & Bränberg, 2015). Although the frequency estimation assumptions are practically impossible to evaluate (Holland, Sinharay, von Davier, & Han, 2008), studies have shown that, with a single anchor, frequency estimation produces more bias than chained equipercentile when group differences are large (e.g., Powers & Kolen, 2014; Wang, Lee, Brennan, & Kolen, 2008). With multiple common variables, these assumptions may be even more difficult to support, yet multiple common variables may improve equating accuracy in the presence of group differences, especially at large sample sizes. Tucker and simplified versions of it are expected to outperform frequency estimation at smaller sample sizes. However, because of their stronger assumptions, they are expected to produce more bias at larger sample sizes, relative to frequency estimation, as is the case with single-anchor equating.

The main research question examined in this study is, how does equating accuracy differ by method over variations in examinee population, sample size, anchor test length, and anchor type? It is anticipated that the benefits of incorporating external covariates, in terms of improved accuracy, will be most evident at smaller anchor test lengths, where anchors are less reliable, less strongly related to total scores, and less able to control for group nonequivalence. The benefits are also expected to be more evident for external versus internal anchor tests. Equating over different sample sizes will then shed light on the numbers of examinees needed to support the more complex estimation required by the multivariate methods, especially frequency estimation. Furthermore, equating over different subsets of a target population will shed light on how accuracy is impacted by differences in group ability. These issues are examined by comparing multivariate methods with EG and NEAT methods with a single common variable within two resampling studies, where pseudo forms of a state science test are equated using internal and external science anchors and linked using reading and math scores as additional external covariates.

Method

Data

The equating methods described above were compared within two resampling studies using data from the 2013/2014 administration of an end-of-year state test. Scored item responses were obtained in science, reading, and math for all general education eleventh graders tested in the state (N = 20,308, after removing examinees with missing data). The science and math tests contained 60 items each, and the reading test contained 50 items, with each item worth one point. Table 1 contains descriptive statistics for each test. Distributions were slightly negatively skewed, with some ceiling effects occurring at the top of each score scale. Reading and math total scores correlated with science at 0.82 and 0.80, respectively, in the full population T. Disattenuated correlations between science and reading, and science and math were 0.90 and 0.86, respectively, with internal consistency

reliabilities (coefficient alpha) of 0.91, 0.92, and 0.94 for science, reading, and math.

The full population of examinees was divided into two pseudo populations, P and Q. Cumulative grade point average (GPA) was used as a population selection variable, with Q consisting of a randomly selected 10,154 examinees having GPA below 3.8 and P consisting of the remaining examinees with unrestricted GPA. As a result, Q had a mean total science score about 3 points lower than P (38.79 compared with 41.88 out of 60). Descriptives for P and Q are also contained in Table 1. Overall, score distributions were slightly less negatively skewed in Q than in P. The correlations between science, reading, and math total scores were slightly stronger in P than in T and Q.

Four pseudo test conditions with varying anchor test lengths were examined in the resampling studies described below. These lengths included $K_V = 8$, 12, 16, and 20 anchor items (labeled $V8$ to $V20$). The shortest anchor test condition, with only 8 items, represents what some testing programs may hope to achieve, given practical constraints on test length and concerns around test security with the increased exposure of anchor items across administrations (discussed by Fitzpatrick, 2008). The longest anchor condition, with 20 items, represents a more ideal length for the NEAT design (e.g., Angoff, 1984).

At each study replication, the science items were randomly assigned to test forms X, Y, and V, while accounting for the science content domain of each item.[1] At each replication, 20 unique items for X and for Y were selected first, balancing by content domain across forms. K_V items were then selected for V. Given the uneven number of items per content domain, V was similar but not always identical in content to X and Y. With internal anchor items, the total test lengths were $K_X = K_Y = 20 + K_V$, depending on the anchor condition. With external anchors, test lengths were always $K_X = K_Y = 20$. Note that other studies have used fixed pseudo tests to examine equating accuracy (e.g., Holland et al., 2008; Sinharay & Holland, 2007). The present study is unique in its use of random sampling to select pseudo forms by replication, discussed further below.

Two preliminary analyses were conducted to examine: a) the linear relationships

among the various pseudo tests, which were found to align with expectations and b) invariance in the population linking and equating functions by gender, which was found to be acceptable. See the Appendix for details.

Resampling Studies

Equating accuracy was examined here using resampling procedures that involved repeated random sampling from T, P, and Q. Resampling allows for the simulation of random variability and estimation of statistical precision without reliance on parametric population models (Efron, 1981). Resampling was conducted under two separate studies, each based on a different approach for sampling from T. In Study 1, random samples of sizes N = 1,000, 2,000, 5,000, and 10,000 were drawn without replacement directly from T at each replication. Study 1 approximated an EG design, where groups were expected to be randomly equivalent, and the need for statistical control via anchor tests or covariates would be minimized. In Study 2, population T was divided into P and Q as described above, so that P scored roughly 3 out of 60 points higher on the full science test than Q. As a result, Study 2 approximated a design with nonequivalent groups, where anchors and covariates could lead to improvements in equating accuracy. Sample sizes in Study 2 matched those of Study 1.

As noted above, forms X, Y, and V were created via random sampling without replacement at each replication of the studies. The unique items per form were always a randomly selected 20, balanced by content, and the number of common items varied by anchor length condition. This random sampling by item ensured that, overall, the pseudo forms were fully representative of the content and statistical properties of the total test. At a given replication, the forms may have differed in statistical characteristics, including difficulty. However, overall, the need for equating is expected to have been minimized and unintentional, as is ideally the case in practice. Averaging across replications then provides a more comprehensive summary of equating accuracy than could be obtained using fixed

pseudo forms which are less likely to represent all the complex relationships among items (for a discussion of related issues, see Michaelides & Haertel, 2014).

Prior to conducting the resampling studies, another preliminary analysis examined the expected form difficulty difference in advance by sampling pseudo forms X and Y, each of length 20 items, 1,000 times from the distribution of 60 science item difficulties obtained in T. Form difficulty was approximated by the sum of item difficulties. The mean form difficulty difference was found to be 0.68 out of 20 points. The largest possible form difficulty difference would be obtained when the easiest 20 items were assigned to one form, e.g., X, and the most difficult to the other, Y. In this case, X would have a mean of 10.69 and Y would have a mean of 16.28, with the mean difference being 5.59 points out of 20. Although such an extreme form difficulty difference would be rare in the resampling studies, a constraint was added to the item selection process so that pseudo forms always had difficulty differences of 2 points or less out of 20.

At each replication of the resampling studies, equating was performed under four designs: (1) an EG equating design with no anchor or covariates, (2) a NEAT equating design with only the science anchor, (3) a NEC linking design with only the reading test as the common variable, and (4) a NEATC linking design with the science anchor supplemented by the reading test. Additional designs incorporating math scores as a second external covariate were also examined. Results were nearly equivalent with the designs including reading scores, and are not reported here. As noted above, equatings were conducted both with the science items in V as an internal anchor, contributing to the total score, and as an external anchor. Note that in the EG design with an internal anchor, X and Y contain common anchor scores but they are not used in equating. Such a design would not be used in practice, and it is included here for completeness. In the EG design with an external anchor, the external anchor is ignored. As a result, this condition represents the typical EG design wherein X and Y contain only unique items and an anchor test is not used.

Under the EG design, identity, linear, equipercentile, and presmoothed equipercentile equating were all performed. Under the NEAT, NEC, and NEATC designs, Tucker linear and frequency estimation equipercentile equating with presmoothing were performed. In the NEAT and NEC designs, chained equipercentile equating was also performed. Loglinear presmoothing (Holland & Thayer, 2000) was performed prior to estimating all equipercentile functions. Loglinear models increased in complexity, with terms for preserving up to six moments for each univariate distribution (science total, science anchor, reading, math), up to two additional moments for each bivariate distribution, up to two moments for each trivariate distribution, and up to two moments for the full multivariate distribution of all four variables. Less complex models were also fit, and the final model chosen at a given replication was the one having the smallest AIC.[2]

Single-Group Criterion and Error

With the pseudo test forms differing by replication, there was no single true criterion equating function. Instead, a criterion was obtained via SG equipercentile equating at each replication, using all examinees in T for the given forms. Presmoothing for the criterion equating always preserved the first six univariate moments in X and Y, and the first two bivariate moments.

The crossing of four anchor test lengths with four sample sizes and two anchor types, internal vs external, resulted in a total of 32 conditions per resampling study, with the criterion equating and estimated linkings and equatings under the four designs conducted at each replication. Smoothing, linking, and equating were performed using the *equate* package (version 2.0-5; Albano, 2016b) in *R* (version 3.3.1; R Core Team, 2016).[3]

Equating accuracy and error are typically defined for a fixed criterion equating function that specifies the true equated score for each raw score in a scale. In the resampling studies, equating accuracy and error could not be defined in the usual way (e.g., Kolen & Brennan, 2014, p. 248). With examinees and test forms both treated as

random effects, and without multiple estimates of any single true equated score, there was no way to partition total equating error into its systematic and random components. Instead, an overall root mean square error was obtained at each replication r as

$$RMSE_r = \sqrt{\frac{1}{K+1}\sum_{i=0}^{K} p_{ri}(\hat{e}_{ir} - e_{ir})^2}, \tag{1}$$

where raw scores range from $i = 0$ to the number of items K, p_{ri} is a weight set to 1, $\hat{e}_{ir}$ is an estimated equated score, and e_{ir} is the true equated score for raw score i at replication r. Means, denoted *MRMSE*, were obtained for each condition and method by averaging over the $R = 1,000$ replications. Weighted means were also obtained for each condition and method by setting p_{ri} to be the smoothed proportion of examinees in T at replication r having raw X score i. Finally, mean errors for the internal anchor conditions were multiplied by the corresponding rescaling constants $21/(K_V + 21)$ so as to adjust for the differing scale lengths and express the results in terms of a 21-point scale, like in the external anchor conditions. In the interest of space, weighted means are not presented here but can be made available upon request. Results differed only slightly for weighted and unweighted mean error.

Summary

The resampling process utilized in Studies 1 and 2 can be summarized in the following steps.

1. Select examinees by randomly sampling either $N \times 2$ examinees from T, in Study 1, or N examinees each from P and Q, in Study 2.

2. Build test forms by randomly sampling, from the 60 science items, 20 unique items each for forms X and Y and K_V common anchor items.

3. Subset the full population distribution for T across all 60 items based on the examinees selected in step 1, and the subsets of items selected in step 2. Calculate

total and anchor scores based on internal and external anchor tests.

4. Find the reading and math total scores for selected examinees, combine with science total scores, and convert the resulting data sets into multivariate frequency distributions for X and Y.

5. Fit to each multivariate distribution a series of nested loglinear models. Compare model fit, select the final model, and apply it to obtain smoothed frequency distributions for X and Y.

6. Estimate the different linking and equating functions, based on their corresponding designs and with internal and external anchor tests, as described above.

7. Find the presmoothed equipercentile criterion equating function based on an SG design that includes all examinees in T taking both X and Y, as determined in step 2.

8. Estimate error for each function based on the discrepancy between the estimates obtained in step 6 and the criterion from step 7.

These eight steps were replicated 1,000 times for Studies 1 and 2. Results were then summarized across replications for a given study, condition, and equating method using mean error.

Results

Results based on *MRMSE* are presented first for Study 1 and then for Study 2. Accuracy is summarized across all sample sizes and equating methods, with a focus on anchor lengths $V8$ and $V20$, which represent the two extremes of the anchor conditions.[4] As noted above, results are excluded for NEATC designs that included both reading and math scores as external covariates, since they did not differ noticeably from results for the NEATC design with reading scores and without math.

Study 1: Resampling from T

Table 2 contains *MRMSE* for Study 1, with results for internal anchors in the top half of the table and external anchor tests in the bottom half. Designs and methods are shown in rows, with anchor lengths ($V8$ and $V20$) and then sample sizes (1,000, 2,000, 5,000, and 10,000) in columns. Note that results for a given column in Table 2, and given columns in subsequent tables, are all based on the same 1,000 replications within the respective resampling study. The smallest error within each column is bolded.

With the internal anchor test, the largest *MRMSE* was 0.50, for identity equating in $V8$, sample sizes 1,000 and 10,000. The smallest *MRMSE* was 0.08 for frequency estimation equipercentile equating under the NEAT design in $V20$, sample size 10,000. Equipercentile equating always produced the smallest error under the EG design, except in $V20$, sample size 1,000, where error was smaller for identity equating. Comparing methods within a given common variable design (NEAT, NEC, and NEATC), anchor condition, and sample size, frequency estimation always produced the smallest error. For a given anchor condition and sample size, comparing across designs, error was smallest in $V8$, sample sizes 1,000 and 2,000 for frequency estimation under the NEATC design. In $V8$, sample sizes 5,000 and 10,000, *MRMSE* was smallest for frequency estimation under both the NEAT and NEATC designs. In $V20$, error was smallest for frequency estimation under the NEAT design, at all sample sizes.

Trends in *MRMSE* with the external anchor test were similar to those found with the internal anchor. Identity equating again produced the largest *MRMSE* of 0.68 in $V8$ at sample sizes 1,000 and 10,000. Errors for identity were close to the estimated form difficulty difference of 0.68 noted above. The smallest *MRMSE* was 0.08 for frequency estimation linking under the NEATC design in $V20$, sample size 10,000. Under the EG design, error was always smallest for a given anchor condition and sample size for equipercentile equating. Comparing methods within a given common variable design, anchor condition, and sample size, error was again always smallest for frequency estimation. Finally, for a

given anchor condition and sample size, comparing across designs, errors tended to be similarly small for frequency estimation under the NEAT, NEC, and NEATC designs. In some of these comparisons frequency estimation under the NEATC design was smallest.

Study 2: Resampling from P and Q

Table 3 contains *MRMSE* for Study 2, both internal and external anchors, and is structured like the previous results tables. With the internal anchor test, the largest *MRMSE* was 1.05, for linear equating in $V8$, sample size 1,000. The smallest *MRMSE* was 0.13 for chained equipercentile equating under the NEC design in $V8$, sample size 10,000. Under the EG design, identity equating always produced smaller error than linear and equipercentile. Under the NEAT design, chained equipercentile equating produced smaller error than Tucker linear and frequency estimation, except in $V20$, sample size 1,000 where it was comparable with frequency estimation. Comparing across designs and methods for a given anchor length and sample size, frequency estimation under the NEATC design produced the smallest errors in $V8$, sample sizes 1,000, 2,000, and 5,000. In $V20$, sample size 1,000, errors were similarly small for Tucker and frequency estimation under the NEATC design. Errors in $V20$, sample sizes 2,000 and 5,000 were smallest for chained NEAT equating, and in $V20$, sample size 10,000, they were similarly small for chained NEAT and NEC equating.

With the external anchor test, *MRMSE* for identity equating again hovered around the expected form difficulty difference of 0.68, as in Study 1. Under the NEAT design, chained equipercentile again tended to produce smaller error than Tucker linear and frequency estimation, with the only exception being in $V20$, sample size 1,000. The smallest *MRMSE* for a given anchor condition and sample size were nearly always produced by frequency estimation under the NEATC design. The only exception was in $V8$, sample size 10,000, where chained NEC equating was similarly small. The smallest *MRMSE* overall was 0.13 for frequency estimation linking under the NEATC design in

$V20$, sample size 10,000.

Discussion

Previous research has shown that a single anchor test can sometimes be insufficient when equating with nonequivalent groups (Moses et al., 2011). Potentially problematic situations include large form difficulty differences, large ability differences between groups, and weak anchor tests. Whereas studies have demonstrated the use of equating with multiple anchors and linking within NEC and NEATC designs, research has not addressed the conditions under which these methods are expected to perform well. Furthermore, simulation and resampling studies, wherein true equating functions are available, have been limited to small sample methods (e.g., Bränberg & Wiberg, 2011; Kim et al., 2011). The overall accuracy of equipercentile methods with multiple common variables has not been thoroughly examined.

This study examined equating accuracy for observed-score linking and equating methods that utilize no common variables (under the EG design), one common variable (the NEAT and NEC designs), and multiple common variables (the NEATC design), over conditions of varying sample size (1,000 through 10,000), anchor test length ($V8$ to $V20$), anchor type (internal versus external), and student ability (population T in Study 1 versus P and Q in Study 2). Resampling of items and examinees using population data from a state testing program resulted in overall estimates of accuracy, via mean $RMSE$, for each method by condition. Results are summarized here by study and discussed in terms of practical applications and future research.

Study 1, with examinees sampled randomly from the full population T, resembled a testing program wherein ability differences between groups taking forms X and Y are expected to be small and random. In terms of mean error, equating was always more accurate than not equating at all. Results showed that some form of equating was always preferable to identity equating. In the absence of a common variable, equipercentile

equating was always most accurate under the EG design. Furthermore, linking under the NEC design was always more accurate than equating without a common variable under the EG design (confirming Wiberg & Bränberg, 2015). Frequency estimation was nearly always the most accurate method, with the smallest errors tending to come from the NEATC design (similar to Wiberg & Bränberg, 2015, who used kernel methods).

Although mean error was minimized under the NEAT, NEC, and NEATC designs in Study 1, the gains were relatively small when compared with results from the EG design. For example, the smallest *MRMSE* in $V8$ was 0.09 for both anchor types, whereas the smallest *MRMSE* under the EG design was 0.11 for the internal anchor and 0.10 for the external anchor (Table 2). In $V20$, the smallest error overall was similarly close to the smallest error under the EG design. Furthermore, increasing the anchor test length had little relative impact on accuracy, with mean errors only decreasing slightly from $V8$ to $V20$ for a given sample size and method. The largest gains in accuracy occurred for larger sample sizes. These findings suggest that the *RMSE* in Study 1 consisted primarily of random error, with minimal bias, as would be expected when groups are randomly equivalent and the assumptions of the equating methods are expected to be met.

Study 2, with examinees sampled randomly from two different subpopulations P and Q, imitated a testing program wherein groups taking X and Y are expected to differ systematically in ability. The need for equating, overall, should be the same as in Study 1, since test forms were generated by replication in the same way. However, the need for common variables should be greater in Study 2, as group differences were systematically larger and thus more likely to confound form difficulty estimates. *MRMSE* for identity equating were the same in Study 2 as in Study 1 (Table 3). Otherwise, mean errors in Study 2 tended to be larger, especially under the EG design. Results from Study 2 showed that identity equating was always preferable to linear or equipercentile under the EG design. This was expected, given the size of the overall ability difference (roughly 3 points out of 60, or 1 point out of 20) compared to the expected average form difficulty difference

(0.68 out of 20 points). Most of the common variable linking and equating methods (NEAT, NEC, and NEATC) were preferable to identity equating for a given condition.

Increases in sample size in Study 2 still produced smaller mean error, as in Study 1. However, the increase in anchor length had more of an impact in Study 2 than 1, with *MRMSE* sometimes decreasing by 0.10 or more from $V8$ to $V20$ for a given design, method, and sample size. For example, with an external anchor, NEAT Tucker linear *MRMSE* at sample size 1,000 decreased by 0.10 from $V8$ to $V20$, as did NEAT frequency estimation (Table 3). In Study 1, the respective decreases were 0.01 and 0.02 (Table 2). Note that, for this same comparison, *MRMSE* for frequency estimation under the NEATC design decreased by 0.01 in Study 1 and did not decrease in Study 2. Similar trends were evident for other sample size comparisons and with internal anchors. These comparisons confirm that common variables are more important with nonequivalent groups. They also indicate that increases in anchor length have less of an impact when the anchor is supplemented by an external covariate. In fact, for the condition presented here (external anchor, sample size 1,000), the smallest mean error in $V8$ was the same as the smallest mean error in $V20$ (0.35). Thus, in terms of mean error, the covariate was able to compensate for an anchor test that was less than half as long.

In Study 2, NEATC frequency estimation again tended to be the most accurate method, especially in $V8$ and at medium to large sample sizes, for internal and external anchors. In $V20$ and at larger sample sizes, chained NEAT and NEC equating sometimes produced comparable or smaller mean errors. The emergence of chained equipercentile linking and equating as the preferable method in some conditions of Study 2, with comparable or larger errors for NEATC frequency estimation, shows that the benefits of chained equating (less bias with large group differences, e.g., Powers & Kolen, 2014) can outweigh the benefits of multiple common variables.

Future research on linking and equating with multiple common variables should consider the following issues. First, the pseudo test forms examined here, with anchors

ranging in length from 8 to 20 items, may be shorter than those used in some large scale testing programs. As test length and anchor test length increase, and the relationship between them is strengthened, the capacity of the anchor test to control for group nonequivalence should increase, thus reducing the need for additional external covariates. The findings of this study are most relevant to testing programs involving tests of small to medium length, and potentially weak anchors. Future studies should examine the benefits and feasibility of NEATC linking with longer total and anchor tests. Second, the testing program used in this study provided easy access to scores on three strongly correlated tests. In practice, high-quality external covariates may not be available. Future research should thus also examine the extent to which the benefits of linking with external covariates depend on the relationships between covariates and total scores. Third, random sampling of items by replication resulted in pseudo forms that, overall, were balanced in content and statistical characteristics. This was considered a novel and useful approach to form creation. However, the pseudo forms did not match any predetermined difficulty level, and did not present a strong need for equating. They also resulted in overall estimates of error, in terms of *RMSE*, that could not be partitioned into systematic and random components. To address these issues, future work should explore the accuracy of these methods using resampling or simulation studies with fixed test forms and/or forms of known difficulty. It is expected that the addition of supplemental common variables in the NEATC design may reduced bias while also increasing standard error. Finally, linking under the NEC and NEATC designs may produce score conversions wherein population invariance does not hold. Future studies should investigate further the trade-off between improvements in accuracy and the loss of population invariance for NEC and NEATC linking methods.

Linking with supplemental external covariates was shown here to produce more accurate results than other methods under a variety of conditions, and it is recommended for consideration whenever external variables are available. However, the use of external variables seems especially appropriate in two specific contexts. The first is in testing

programs not designed or not well suited for equating with an internal anchor test (e.g., Wiberg & Bränberg, 2015), including testing programs that utilize an EG design with groups that may not be randomly equivalent (e.g., Lyren & Hambleton, 2011). The second is in less traditional testing programs, such as those involving small-scale, formative, or longitudinal assessments. In these cases, tests are often brief and internal anchors may be limited or nonexistent, but external measures may be readily available. In either context, findings suggest that sample sizes of 1,000 are sufficient to realize the benefits of incorporating external variables. Additional research, including real data applications, is needed to further clarify the benefits of linking and equating with multiple common variables in these contexts.

References

Albano, A. D. (2016a). equate: an R package for observed-score linking and equating. *Journal of Statistical Software*, *74*(8), 1–36.

Albano, A. D. (2016b). equate: observed-score linking and equating. Retrieved from http://CRAN.R-project.org/package=equate.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: ETS.

Bränberg, K. & Wiberg, M. (2011). Observed score linear equating with covariates. *Journal of Educational Measurement*, *48*, 419–440.

Cook, L. L. & Paterson, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement*, *11*, 225–244.

Dorans, N. J. & Holland, P. W. (2000). Population invariance and the equatability of tests: basic theory and the linear case. *Journal of Educational Measurement*, *37*, 281–306.

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika*, *68*(3), 589–599.

Fitzpatrick, A. R. (2008). Ncme 2008 presidential address: The impact of anchor test configuration on student proficiency rates. *Educational Measurement: Issues and Practice*, *27*(4), 34–40.

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.

Holland, P. W., Sinharay, S., von Davier, A. A., & Han, N. (2008). An approach to evaluating the missing data assumptions of the chain and post-stratification equating methods for the NEAT design. *Journal of Educational Measurement*, *45*, 17–43.

Holland, P. W. & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, *25*, 133–183.

Kim, S., Livingston, S. A., & Lewis, C. (2011). Collateral information for equating small samples: a preliminary investigation. *Applied Measurement in Education*, *24*, 302–323.

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

Kolen, M. J. & Brennan, R. L. (2014). *Test equating, scaling, and linking: methods and practices*. New York, NY: Springer.

Liou, M., Cheng, P. E., & Li, M. (2001). Estimating comparable scores using surrogate variables. *Applied Psychological Measurement*, *25*, 197–207.

Liu, J., Sinharay, S., Holland, P. W., Curley, E., & Feigenbaum, M. (2011). Test score equating using a mini-version anchor and a midi anchor: a case study using SAT® data. *Journal of Educational Measurement*, *48*, 361–379.

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

Lyren, P.-E. & Hambleton, R. K. (2011). Consequences of violated equating assumptions under the equivalent groups design. *International Journal of Testing*, *11*(4), 308–323.

Michaelides, M. P. & Haertel, E. H. (2014). Selection of common items as an unrecognized source of variability in test equating: A bootstrap approximation assuming random sampling of common items. *Applied Measurement in Education*, *27*, 46–57.

Moses, T., Deng, W., & Zhang, Y. (2011). Two approaches for using multiple anchors in NEAT equating: a description and demonstration. *Applied Psychological Measurement*, *35*, 362–379.

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

R Core Team. (2016). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org/

Sansivieri, V. & Wiberg, M. (2017). IRT observed-score equating with the non-equivalent groups with covariates design. In L. A. van der Ark, M. Wiberg, S. A. Culpepper,

J. A. Douglas, & W.-C. Wang (Eds.), *Quantitative psychology - 81st annual meeting of the psychometric society* (pp. 275–285). New York, NY: Springer.

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*, 249–275.

von Davier, A. A. (Ed.). (2011). *Statistical models for test equating, scaling, and linking.* New York, NY: Springer.

Wallin, G. & Wiberg, M. (2017). Non-equivalent groups with covariates design using propensity scores for kernel equating. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.-C. Wang (Eds.), *Quantitative psychology - 81st annual meeting of the psychometric society* (pp. 309–320). New York, NY: Springer.

Wang, T., Lee, W., Brennan, R. L., & Kolen, M. J. (2008). A comparison of the frequency estimation and chained equipercentile methods under the common-item non-equivalent groups design. *Applied Psychological Measurement*, *32*, 632–651.

Wiberg, M. & Bränberg, K. (2015). Kernel equating under the non-equivalent groups with covariates design. *Applied Psychological Measurement.*

Footnotes

[1]The four content domains were: (1) inquiry, the nature of science, and technology, with 13 items; (2) physical science, with 16 items; (3) life science with 17 items; and (4) earth and space science, with 14 items.

[2]Model fit was also compared using chi-square likelihood ratio tests, but equating accuracy did not differ noticeably from results based on AIC model fit.

[3]Example *R* code demonstrating the programming behind the resampling studies can be obtained from the corresponding author.

[4]Results for $V12$ and $V16$, which followed the trend evident in $V8$ to $V20$, can be obtained from the corresponding author.

Table 1
Total Score Descriptive Statistics for Population Distributions

					Correlation	
Pop	Variable	*Mean*	*SD*	*Skew*	Science	Reading
T	Science	40.34	10.75	-0.49		
	Reading	34.62	9.86	-0.64	0.82	
	Math	39.63	12.84	-0.31	0.80	0.77
P	Science	41.88	10.84	-0.61		
	Reading	36.03	9.84	-0.78	0.83	
	Math	41.78	13.12	-0.49	0.82	0.79
Q	Science	38.79	10.44	-0.42		
	Reading	33.21	9.68	-0.55	0.80	
	Math	37.48	12.19	-0.19	0.76	0.73

Note: Pop refers to the student population, where the total population T contained all 20,308 students, and pseudo populations P and Q contained 10,154 students each.

Table 2
Study 1 MRMSE *for Internal and External Anchors of Lengths 8 and 20 Over Designs, Methods, and Sample Sizes*

Internal anchor

		$V8$				$V20$			
Design	Method	1,000	2,000	5,000	10,000	1,000	2,000	5,000	10,000
EG	I	0.50	0.48	0.49	0.50	0.35	0.35	0.35	0.35
	L	0.45	0.40	0.38	0.38	0.39	0.35	0.31	0.30
	E	0.38	0.28	0.17	0.11	0.39	0.29	0.17	0.11
NEAT	LT	0.40	0.38	0.37	0.38	0.31	0.30	0.29	0.29
	EC	0.35	0.25	0.15	0.10	0.31	0.24	0.14	0.09
	EF	0.32	0.23	**0.14**	**0.09**	**0.28**	**0.21**	**0.12**	**0.08**
NEC	LT	0.41	0.38	0.37	0.37	0.35	0.33	0.30	0.29
	EC	0.43	0.32	0.20	0.14	0.43	0.32	0.21	0.14
	EF	0.34	0.24	0.15	0.10	0.33	0.25	0.15	0.10
NEATC	LT	0.40	0.37	0.37	0.37	0.31	0.30	0.29	0.29
	EF	**0.31**	**0.22**	**0.14**	**0.09**	0.29	0.23	0.15	0.10

External anchor

		$V8$				$V20$			
Design	Method	1,000	2,000	5,000	10,000	1,000	2,000	5,000	10,000
EG	I	0.68	0.65	0.67	0.68	0.66	0.66	0.66	0.65
	L	0.53	0.49	0.47	0.47	0.53	0.51	0.47	0.46
	E	0.35	0.26	0.17	0.10	0.35	0.26	0.16	0.10
NEAT	LT	0.50	0.48	0.47	0.47	0.49	0.48	0.47	0.46
	EC	0.40	0.29	0.18	0.11	0.41	0.31	0.20	0.13
	EF	0.34	0.24	0.15	**0.09**	**0.32**	0.24	**0.14**	0.09
NEC	LT	0.49	0.47	0.46	0.47	0.50	0.49	0.47	0.45
	EC	0.43	0.30	0.20	0.13	0.43	0.31	0.20	0.14
	EF	0.34	**0.23**	0.15	**0.09**	0.33	0.24	**0.14**	0.09
NEATC	LT	0.49	0.47	0.46	0.47	0.49	0.48	0.47	0.45
	EF	**0.33**	**0.23**	**0.14**	**0.09**	**0.32**	**0.23**	**0.14**	**0.08**

Note: Methods are abbreviated with an initial I for identity, L for linear, and E for equipercentile, and a subsequent T for Tucker, C for chained, and F for frequency estimation. The smallest values in each column are bolded.

Table 3
Study 2 MRMSE *for Internal and External Anchors of Lengths 8 and 20 Over Designs, Methods, and Sample Sizes*

Internal anchor

		$V8$				$V20$			
Design	Method	1,000	2,000	5,000	10,000	1,000	2,000	5,000	10,000
EG	I	0.50	0.48	0.49	0.50	0.35	0.35	0.35	0.35
	L	1.05	1.03	1.02	1.02	1.01	1.01	1.00	1.00
	E	0.92	0.89	0.86	0.85	0.94	0.89	0.86	0.85
NEAT	LT	0.50	0.47	0.47	0.47	0.33	0.33	0.32	0.31
	EC	0.40	0.32	0.24	0.21	0.34	**0.25**	**0.18**	**0.14**
	EF	0.43	0.37	0.32	0.30	0.33	0.26	0.20	0.18
NEC	LT	0.45	0.41	0.41	0.41	0.38	0.36	0.35	0.34
	EC	0.44	0.33	0.20	**0.13**	0.45	0.32	0.20	**0.14**
	EF	0.39	0.31	0.23	0.20	0.40	0.32	0.25	0.23
NEATC	LT	0.42	0.39	0.39	0.39	**0.32**	0.31	0.30	0.30
	EF	**0.34**	**0.26**	**0.18**	0.15	**0.32**	0.26	0.20	0.18

External anchor

		$V8$				$V20$			
Design	Method	1,000	2,000	5,000	10,000	1,000	2,000	5,000	10,000
EG	I	0.68	0.65	0.67	0.68	0.66	0.66	0.66	0.65
	L	1.09	1.07	1.07	1.07	1.08	1.10	1.08	1.08
	E	0.91	0.88	0.86	0.85	0.91	0.88	0.85	0.85
NEAT	LT	0.65	0.62	0.62	0.63	0.55	0.55	0.53	0.52
	EC	0.43	0.34	0.24	0.20	0.42	0.32	0.21	0.14
	EF	0.51	0.46	0.41	0.39	0.41	0.33	0.27	0.24
NEC	LT	0.53	0.49	0.49	0.50	0.52	0.51	0.50	0.49
	EC	0.44	0.34	0.22	**0.16**	0.46	0.34	0.22	0.15
	EF	0.37	0.30	0.22	0.19	0.38	0.30	0.22	0.19
NEATC	LT	0.51	0.47	0.48	0.48	0.50	0.49	0.47	0.46
	EF	**0.35**	**0.28**	**0.19**	**0.16**	**0.35**	**0.26**	**0.18**	**0.13**

Note: Methods are abbreviated with an initial I for identity, L for linear, and E for equipercentile, and a subsequent T for Tucker, C for chained, and F for frequency estimation. The smallest values in each column are bolded.

Appendix

A preliminary analysis was conducted to examine the linear relationships among the various pseudo tests. Mean correlations between total scores (including the internal anchor) and anchor, reading, and math scores were estimated over 500 randomly sampled pseudo tests for each anchor length. For the full population T, correlations between anchor and total scores increased from 0.83 at $V8$ to 0.94 at $V20$. Correlations for reading and math increased from 0.78 and 0.76 at $V8$, to 0.80 and 0.78 at $V20$. Correlations for P and Q differed only slightly from those for T for a given pair of variables. Over these same 500 samples, mean internal consistency reliabilities were also obtained by anchor length condition, for the total and anchor tests. Reliabilities increased from 0.82 to 0.87 for the total tests, and from 0.56 to 0.76 for the anchors, from $V8$ to $V20$. Reliabilities increased similarly in P and Q, but were slightly lower overall for population Q compared to P.

A second preliminary analysis examined invariance in the population linking and equating functions by gender. A set of pseudo forms was created by randomly assigning the 60 science items to X, Y, and V, while controlling for content. Each form contained the maximum 20 items. Equipercentile equating under an SG design was used to equate X to Y, first in the full population T, then again with all female examinees in T, and all male examinees in T. The unstandardized root expected mean square difference (Dorans & Holland, 2000), which measures the discrepancy in equating by subpopulation from the full population, was found to be 0.07. Next, X and Y were equated under a nonequivalent groups design using P and Q, again separately by gender. Form V, treated as an external anchor, was supplemented by reading and math total scores. The discrepancy in multivariate frequency estimation equating for female and male examinees relative to SG equating in T, again obtained as the unstandardized root expected mean square difference, was 0.08. These results indicate that the average discrepancy in equating by subpopulation should be negligible, below a tenth of a score point, and incorporating external covariates

with nonequivalent groups should not substantially impact population invariance.