IAEA Conference 2019

The Influence of Unbalanced Group Sizes on the Choice of Equating Methods Under the Nonequivalent Groups Anchor Test (NEAT) Design: A Monte Carlo Simulation Study

Leina Zhu[1], Yvette Song*[2] and Fei Wang[3]

[1, 2]OTIEA (ONETARGET Institute for Educational Assessment), Beijing, China, 100102

[3]University of International Business and Economics, Beijing, 100029, China

Abstract

ONETARGET Institute for Educational Assessment (OTIEA) is a research-oriented assessment service to schools, as well as directly to individuals in mainland China. Each year, OTIEA conducts assessments to thousands of students from 1-12 on students' cognitive ability, motivation, self-regulation, and environmental impact using affluent scales. Frequently, students' scores on these scales are compared across different administrations or grades. It is noted that equating should be implemented to ensure that students' scores of different administrations and grades are comparable to each other. Although various equating methods have been proposed regarding different equating situations, significant issues in techniques and applications of equating in practice need our continuous attention.

The purpose of this monte carlo simulation study is to investigate the choice of equating methods under the nonequivalent groups anchor test (NEAT) design and unbalanced/unequal group sizes. In the NEAT design, group mean differences and variances were attributed to two variations, i.e., test form differences and examinee group differences (Kolen & Brennan, 2004). In previous empirical and simulation studies, the examinee group differences were studied on population ability differences under the NEAT design (e.g., Brennan, 1990; Dorans & Holland, 2000; Hanson, 1991; Kolen & Brennan, 2004; Moses & Kim, 2007). In addition to the ability differences, this study considers educational realistic scenarios of unbalanced group sizes when comparing groups. One of the considerations is that when test scores are discrete (e.g., number-correct scores), some scores could not find equivalent counterparts on equated test forms using observed score equating. This case may be more exacerbating when two groups

*Corresponding author: yvette.song@onetarget.cn

differ in population ability and in addition, have unbalanced group sizes because unequal sample sizes may result in larger variances differences between groups.

Therefore, the purpose of this study is to evaluate the effects of variation in unbalanced sample sizes between equated groups on equating methods under the NEAT design. The equating methods include equipercentile equating methods and linear equating methods. Following the same procedures in González and Wibergs's (2017), the effect of unbalanced group sizes on the choice of equating methods will be investigated in a simulation study across 5 simulation conditions (i.e., 5 unbalanced group sizes). Different equating methods (i.e., linear and equipercentile equating methods) will be compared in terms of standard equating error (SE) of equating, relative bias, and root mean square errors (RMSE). It is hypothesized that variation in unbalanced sample sizes between groups will impact on equating methods under the NEAT design. Implications to researchers and practitioners regarding the choice of method for score equating under the NEAT design and unbalanced group sizes will thereby be discussed.

Introduction

The nonequivalent groups anchor test (NEAT) design is a widely used equating design in test equating. Under the NEAT design, two groups of examinees are administered two test forms which have some items in common. Due to different test forms used, in addition to the group difference in ability levels (i.e., from different population), test form difference is also introduced into the two groups' test scores. Test scores on the two test forms should then be equated through the anchor/common test scores. In general, the anchor/common test is used to adjust for scores differences in ability in two samples (Kolen & Brennan, 2014; Lu & Guo, 2018).

Different equating methods have been developed to deal with data collected through the NEAT design. Two kinds of traditional equating methods are discussed: the equipercentile and the linear equating methods. The linear equating methods include the Tucker methods, the Levine observed-score method, and the chained linear equating. The equipercentile equating methods include the frequency estimation method and the chained equipercentile method. Previous studies have evaluated various equating methods in different circumstances under the NEAT design (e.g., Holland, et al., 2008; Kolen, Brennan, 2014; Moses, Deng, & Zhang, 2011; Moses & Holland, 2010a, 2010b; Mroch, et al., 2009; Puhan & Liang, 2011; Sinharary & Holland, 2010a, 2010b; Suh, et al., 2009; Wang et al., 2008; Zu & Yuan, 2012). For example, von Davier et al. (2004) showed that similar equating results could be obtained while comparing chained equipercentile and frequency estimation equipercentile methods given the two groups' population abilities are similar. Wang et al. claimed that the frequency estimation method resulted in larger bias compared to the chained equipercentile method when group differences were noticeable (Wang et al., 2008). Similarly, studies found that chained equipercentile methods tended to produce less biased equating results relative to frequency estimation method when group differences were substantial (Hagge and Kolen 2011, 2012; Holland et al. 2008; Lee et al. 2012; Liu & Kolen, 2011; Powers et al. 2011; Powers & Kolen, 2011, 2012; Sinharay, 2011; Sinharay and Holland 2010a, 2010b;

Sinharay et al. 2011). Nevertheless, when group differences were neglectable, Tucker and frequency estimation equipercentile equating were preferred over other equating methods (Kolen, Brennan, 2014). Summarizing the previous studies could find that different equating methods produced similar or very different results which depended on certain circumstances.

The pros and cons of each equating method have been discussed under certain circumstances. It was noted that when comparing different equating methods, nearly all studies addressed group differences in population ability, e.g., mean and variances differences in the observed scores (e.g., Brennan, 1990; Dorans & Holland, 2000; Hanson, 1991; Kolen & Brennan, 2014; Moses & Kim, 2007). So far, yet no studies address the issue of unbalanced/unequal group sizes among groups in test equating. In educational realistic scenarios, however, it is not uncommon that we encounter issues of unbalanced/unequal group sizes when comparing groups (e.g., Herrera & Gómez, 2008; Hubbard & Seddon, 1989; Jung & Ahn., 2005; Miller, Morgan, Espeland, & Emerson, 2001). In the present study, we investigated the influence of unbalanced group sizes on the choice of equating methods under the NEAT design. The methods included in this study are linear and equipercentile equating for data collected through the NEAT design.

Linear equating and equipercentile equating methods for the NEAT design

We started by illustrating linear equating methods under the NEAT design. Generally, linear equating method is to equate observed scores on X (test X) to the scale of observed scores on Y (test Y), that is, to convert observed test X scores to the Y scores as follows (Kolen & Brennan, 2014, p. 104):

$$l_{YS}(\mathrm{x}) = \frac{\sigma_s(Y)}{\sigma_s(X)}[X - \mu_s(X)] + \mu_s(Y) \quad (1)$$

where s indicates the synthetic population. According to the NEAT design, the two sample groups come from two populations, which form the synthetic population s. The means and standard deviations in Equation (1) can be expressed as follows:

$$\mu_s(X) = w_1\mu_1(X) + w_2\mu_2(X), \quad (2)$$

$$\mu_s(Y) = w_1\mu_1(Y) + w_2\mu_2(Y), \quad (3)$$

$$\sigma_s^2(X) = w_1\sigma_1^2(X) + w_2\sigma_2^2(X) + w_1w_2[\mu_1(X) - \mu_2(X)]^2, \quad (4)$$

$$\sigma_s^2(Y) = w_1\sigma_1^2(Y) + w_2\sigma_2^2(Y) + w_1w_2[\mu_1(Y) - \mu_2(Y)]^2, \quad (5)$$

where the subscripts 1 and 2 refer to the Populations 1 and 2, respectively. As defined in Braun and Holland's (1982) work, Populations 1 and 2 are weighted by w_1 and w_2, respectively, where $w_1 + w_2 = 1$ and $w_1, w_2 \geq 0$.

Second, we illustrate two possible approaches to perform equipercentile equating under a NEAT design, i.e., frequency estimation and chained equating. Under the NEAT design, the first approach of equipercentile equating methods is frequency estimation. The conditional score distribution of X and Y are defined as $f_{XP}(x|\alpha)$ and $f_{YP}(y|\alpha)$, respectively on population P. Corresponding conditional score distribution of X and Y are defined as $f_{XQ}(x|\alpha)$ and $f_{YQ}(y|\alpha)$, respectively on population Q. The distributions of

anchor test scores are defined as $f_{AP}(\alpha)$ and $f_{AQ}(\alpha)$ on two populations. As defined, T denotes the synthetic population and

$$f_{XT}(x) = w_p f_{XP}(x) + w_Q \sum_a f_{XP}(x|a) f_{AQ}(a), \quad (6)$$

$$f_{XT}(x) = w_p f_{XP}(x) + w_Q \sum_a f_{XP}(x|a) f_{AQ}(a). \quad (7)$$

The second approach of equipercentile equating methods is chained equipercentile equating. In this method, text form X is equated to test form A and then test from A is equated to test form Y. The chained transformation is described as

$$\varphi_Y(x) = F_{YQ}^{-1}(F_{AQ}(F_{AP}^{-1}(F_{XP}(x)))) = \varphi_{YQ}(\varphi_{AP}(x)). \quad (8)$$

Unequal group sizes

Group comparisons (e.g., group means comparisons) are very common in social and behavioral sciences, such as comparing the experimental and control groups whether they are adjacent in their mean levels of abilities. Under unequal sample sizes, variances discrepancies are likely to occur. Taken *t*-test as an example, the sample group with larger sample size is given more weights when calculating pooled variances of the two groups relative to the sample with small sample size. Therefore, the large sample with more cases is more influential in the estimation of the population variance than the small sample. In this case, the condition of unequal sample sizes should be considered when comparing groups. Scenarios include whether the unbalanced sample sizes result in violation of equal variance assumption or the influence of different sample weights assigned in parameter and standard error estimations due to unequal sample sizes. In test equating, when equating scores from two sample groups under the NEAT design, the occurrences of unequal sample sizes should be examined as one group has more cases would be given more weights in estimating the population variance.

Methods

To examine the impact of unbalanced group sizes on the choice of equating methods under the nonequivalent groups anchor test (NEAT) design, we simulated data to mimic the NEAT design. In addition, we focused on generating simulated data with the same experimental conditions except that the two groups have varying sample sizes. We then used different equating methods to analyze the simulated data. In the present study, we considered comparing five equating methods, including the Tucker, Levine, chained linear, frequency estimation, and chained equipercentile equating methods. The impact of unbalanced groups sizes and different equating methods in test equating was evaluated in the set conditions in terms of three criteria, including the bias, standard errors (SE) of equating, and root mean square errors (RMSE). From examining the three criteria, the lowest values for SE, bias, and RMSE were desired indicating more accurate and stable equating.

Following the simulation design of González and Wibergs's (2017), two test forms (both containing 10 items) and an anchor test (also containing 10 items) are administered to

two groups of test takers. One condition was that the two groups each have 1,000 test takers and the other simulation condition was that one group has 1000 test takers whereas the other group varied in the sample sizes, specifically, the other group has far less test takers (e.g., 200, 400, 600, 800, with an interval of 200). In other words, the condition of unequal sample sizes was realized from varying the other group's size while keep one group's size as 1,000 test takers. The item response data were assumed to have been generated by a 2PL model, that is, the guess parameters (c) were set as 0. Given values for discrimination parameters (a) were generated from uniform distributions whereas both item parameters (b) and ability parameters (Θ) were generated from normal distributions. Specifically, distribution of distribution parameter (a) was defined as μ [0.3, 1.5] for both test forms X and Y, in addition, anchor test form A. 10 item difficulty parameters (b) were generated from normal distributions. Similarly, ability parameters (Θ) were generated from normal distributions with mean of 0.5 and 0 for the test form X and Y, respectively.

Due to sampling error, some irregularities often occurred in score distributions. Next, to reduce the influence of sampling error, we presmoothed the score distributions obtained from the NEAT design. Last, the presmoothed data were analyzed using different equating methods as described above. Estimations of equating SE, bias, and RMSE were thereby calculated. To obtain SE, bias, and RMSE, bootstrap function in equating was implemented (Efron 1982; Efron & Tibshirani 1993). Note that because "true" equated values do not exist, equated values estimated from the frequency estimation method were set up as the criterion equated values in the present study. Therefore, equated scores from the frequency estimation method were the criterion equated values during the bootstrap estimation to calculate the SE, bias, and RMSE.

Results

This study evaluated the use of different equating methods for the NEAT data collection designs, in addition, under equal and unequal sample sizes. The advantages and disadvantages of different equating methods was evaluated in terms of bias, standard errors (SE) of equating, and root mean square errors (RMSE). We summarized findings of SE, bias, and RMSE in Table 1 and Figure 1. Next, we discussed the finding with regard to equal sample size condition and unequal sample size condition. First, when the two groups have a balanced sample size in comparison (1000:1000), Tucker outperforms the other methods (see Table 1) in terms of error reduction, with a lowest SE (i.e., 0.10). Mean RMSE was 0.26 and mean RMSE for the remaining methods are between 0.32 (chained) and 0.81 (Levine) except frequency estimation method. As mentioned before, equated values estimated from the frequency estimation method were set up as the criterion. Therefore, the frequency estimation method was the criteria method to obtain the bias and RMSE relative to other equating methods. Meanwhile, visually we could see that Tucker outperforms the other methods (see Figure 1) in terms of SE and bias, with the Tucker curve positioning the lowest location.

Second, with the decreasing sample size (i.e., unequal sample size condition), SE of equating inflated regardless of equating methods (i.e., Tucker, Levine, chained linear,

frequency estimation, or chained equipercentile equating methods). Increasing SE indicated that the equated scores became more biased and demonstrated the impact of unbalanced sample size. More divergence in sample sizes between the two groups would result in more biased equating results. In other words, the equated scores in test equating were less reliable and stable when the sample sizes between equated groups were unbalanced. Among the five methods under the unbalanced sample size design condition, overall, Tucker method outperforms the other methods with the lowest SE values. The changes in bias and RMSE values were not as obvious as in SE when the sample size become more unbalanced.

Overall, Tucker method was more efficient in test equating compared to the other four methods, i.e., Levine, chained linear, frequency estimation, and chained equipercentile equating methods. The unequal sample sizes between equated groups would undermine the efficiency of equating methods. With increasing difference between the sample sizes of the equated groups, the equated scores became more biased and unstable.

Table 1 Standard error, bias, RMSE of different equating methods under various unequal sample sizes

Sample sizes		Tucker	Levine	Chained linear	Frequency estimation	Chained equipercentile
200	SE	0.19	0.52	0.22	0.24	0.28
	Bias	0.21	0.89	0.31	0.02	0.33
	RMSE	0.28	1.02	0.38	0.24	0.44
400	SE	0.14	0.37	0.17	0.18	0.23
	Bias	0.25	0.89	0.32	0.02	0.32
	RMSE	0.29	0.96	0.36	0.18	0.39
600	SE	0.14	0.34	0.16	0.18	0.20
	Bias	0.22	0.85	0.30	0.01	0.32
	RMSE	0.26	0.92	0.34	0.18	0.38
800	SE	0.11	0.30	0.14	0.14	0.18
	Bias	0.20	0.88	0.30	0.03	0.31
	RMSE	0.23	0.93	0.33	0.15	0.36
1000	SE	0.10	0.25	0.13	0.14	0.17
	Bias	0.24	0.77	0.30	0.01	0.31
	RMSE	0.26	0.81	0.32	0.14	0.35

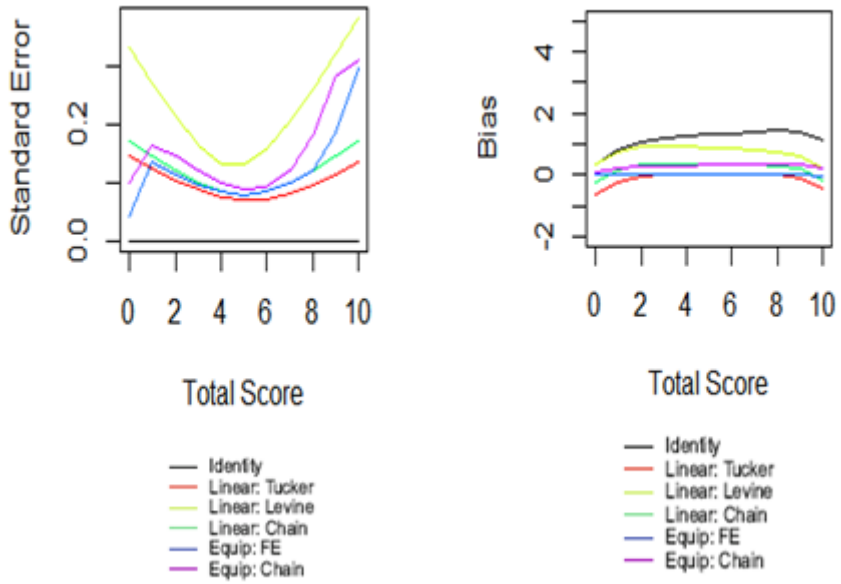

Figure 1 Standard errors and bias for the five equating methods in the equal same size condition (1000:1000)

Discussion

In previous empirical and simulation studies, the examinee group differences were studied only on population ability differences under the NEAT design. No study has ever examined the impact of unbalanced sample sizes between equated groups. The present study intended to evaluate the influence of unbalanced sample size design while using different equating methods. Results from the simulations suggested group size heterogeneity could be an influential factor in test equating. When unequal/unbalanced sample sizes occurred between equated groups, the divergence in sample size resulted in more bias in the equated scores regardless of whichever equating methods researchers chose. Therefore, when implementing test equating, researchers and practitioners should be caution whether equated groups had balanced sample sizes. Among the Tucker, Levine, chained linear, frequency estimation, and chained equipercentile equating methods, Tucker method was more superior producing less biased equating scores than the other four methods. In particular, Tucker method still performed better than the other four methods when the equated groups had unequal sample sizes. Tucker method was thereby recommended as a preferred equating method for data collected from the NEAT design in comparing linear and equipercentile equating methods. Furthermore, when sample sizes were unbalanced, Tucker method still outperformed Levine, chained linear, frequency estimation, and chained equipercentile equating methods.

The same as most simulation studies, results obtained in the simulation study were limited to the study design. These limitations needed further investigations. First, the data generation model specified in this study was a 2PL model. In practice, researchers used

different IRT models (e.g., 1PL, 3PL, etc.). Second, this study had the 2PL model parameter (i.e., a, b, and Θ) values fixed. In real data analysis, model parameters were diversely distributed ranging from small to large values. The third limitation was that the various sample size conditions were limited with few variations (i.e., 1000 vs. 800, 600, etc.) In practice, unbalanced sample sizes would be very different and the impact of unbalanced sample size should be further evaluated case by case. Therefore, further research was needed investigating the impact of various unequal sample sizes, nonequivalent groups in ability levels, different equating methods besides the linear and equipercentile equating methods mentioned in this study.

Nevertheless, with these limitations abovementioned, this study highlighted two important searches: 1) besides focusing on the nonequivalence in ability levels across equated groups, unbalanced sample sizes should be cautioned as well, 2) different equating methods should be evaluated and compared in order to choose the appropriate equating method. This study evaluated the impact of unbalanced sample sizes on choosing different equating methods. It was suggested to use the Tucker method which produced less biased equated scores when implementing test equating between equated groups. This study should be viewed in light of the complexity of conducting test equating across groups and should examine all possible impact with respect to choose the appropriate equating methods.

References

Brennan, R. L. (1990). *Congeneric models and Levine's linear equating procedures* (ACT Research Report 90–12). Iowa City, IA: American College Testing.

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.

González, J., Wiberg, M., Samhällsvetenskapliga fakulteten, Statistik, Umeå universitet, & Handelshögskolan vid Umeå universitet. (2017). *Applying test equating methods : Using R.* Cham: Springer. doi:10.1007/978-3-319-51824-4

Hanson, B. A. (1991). *Method of moments estimates for the four-parameter beta compound binomial model and the calculation of classification consistency indexes*. Iowa City, IA: ACT. (Research Report 91–5)

Hagge, S. L., & Kolen, M. J. (2011). Equating mixed-format tests with format representative and non-representative common items. In M. J. Kolen &W. Lee (Eds.), Mixed-format tests: Psychometric properties with a primary focus on equating (volume 1). (CASMA Monograph Number 2.1) (pp. 95–135). Iowa City, IA: CASMA, The University of Iowa.

Hagge, S. L., & Kolen,M. J. (2012). Effects of group differences on equating using operational and pseudo-tests. In M. J. Kolen & W. Lee (Eds.), Mixed-format tests:

Psychometric properties with a primary focus on equating (volume 2). (CASMA Monograph Number 2.2) (pp. 45–86). Iowa CIty, IA: CASMA, The University of Iowa.

Herrera, Aura-Nidia, & Gómez, Juana. (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel–Haenszel and logistic regression techniques[J]. *Quality & Quantity*, 42(6).

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.

Hubbard J. I., & Seddon, G. M. (1989). Comparison of the Standard and Reliability of the Assessments of Practical Scientific Skills using Groups of Different Sizes[J]. *Research in Science & Technological Education*, 7(1).

Jung, Sin‐Ho, Chul W. Ahn. (2005) Sample size for a two‐group comparison of repeated binary measurements using GEE[J]. *Statistics in Medicine*, 24(17).

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices.* third edition. New York: Springer. doi:10.1007/978-1-4939-0317-7

Lee,W., He, Y., Hagge, S. L.,Wang,W., & Kolen, M. J. (2012). Equating mixed-format tests using dichotomous common items. In M. J. Kolen &W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (volume 2). (CASMA Monograph Number 2.2) (pp.13–44). Iowa CIty, IA: CASMA, The University of Iowa.

Liu, C.,&Kolen, M. J. (2011). A comparison among IRT equatingmethods and traditional equating methods formixed-format tests. In M. J.Kolen&W. Lee (Eds.), *Mixed-format tests:Psychometric properties with a primary focus on equating* (volume 1). (CASMA Monograph Number 2.1) (pp. 75–94). Iowa City, IA: CASMA, The University of Iowa.

Lu, R., & Guo Hongwen (2018). A Simulation Study to Compare Nonequivalent Groups With Anchor Test Equating and Pseudo‐Equivalent Group Linking [J]. *ETS Research Report Series*, 2018(1).

Miller, M E, Morgan T M, Espeland M A, Emerson S S. (2001). Group comparisons involving missing data in clinical trials: a comparison of estimates and power (size) for some simple approaches.[J]. *Statistics in Medicine*, *20*(16).

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.

Moses, T.,& Holland, P.W. (2010). The effects of selection strategies for bivariate loglinear smoothing models on NEAT equating functions. *Journal of Educational Measurement, 47*, 76–91.

Moses, T., & Holland, P. W. (2010b). A comparison of statistical selection strategies for univariate and bivariate log-linear models. *British Journal of Mathematical and Statistical Psychology*, 63, 557–574.

Mroch, A. A., Suh, Y., Kane, M. T., & Ripkey, D. R. (2009). An evaluation of five linear equating methods for the NEAT design. *Measurement, 7*, 174–193.

Powers, S. J., Hagge, S. L., Wang, W., He, Y., Liu, C., & Kolen, M. J. (2011). Effects of group differences on mixed-format equating. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (volume 1). (CASMA Monograph Number 2.1) (pp. 51–73). Iowa City, IA: CASMA, The University of Iowa.

Powers, S. J., & Kolen, M. J. (2011). Evaluating equating accuracy and assumptions for groups that differ in performance. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (volume 1). (CASMA Monograph Number 2.1) (pp. 137–175). Iowa City, IA: CASMA, The University of Iowa.

Puhan, G., & Liang, L. (2011). Equating subscores under the nonequivalent anchor test (NEAT) design. Educational Measurement: Issues and Practice, 30(1), 23–35.

Sinharay, S. (2011). Chain equipercentile equating and frequency estimation equipercentile equating: Comparisons based on real and simulated data. In N. J. Dorans&S. Sinharay (Eds.), *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland.* Lecture Notes in Statistics 202 (pp. 203–219). New York: Springer.

Sinharay, S., & Holland, P. W. (2010a). The missing data assumptions of the NEAT design and their implications for test equating. *Psychometrika, 75*, 309–327.

Sinharay, S., & Holland, P. W. (2010b). A new approach to comparing several equating methods in the context of the NEAT design. *Journal of Educational Measurement, 47*, 261–285.

Sinharay, S., Holland, P. W., & von Davier, A. A. (2011). Evaluating the missing data assumptions of the chain and poststratification equating methods. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 281–296). New York: Springer.

Suh, Y., Mroch, A. A., Kane, M. T., & Ripkey, D. R. (2009). An empirical evaluation of five linear equating methods for the NEAT design. *Measurement, 7*, 147–173.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York: 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.

Zu, J., & Yuan, K.-H. (2012). Standard error of linear observed-score equating for the NEAT design with nonnormally distributed data. *Journal of Educational Measurement, 49*, 190–213.