British Journal of Education
Vol.10, Issue 13, pp.56-67, 2022
Online ISSN: 2054-636X (Online)
Print ISSN: 2054-6351(Print)

Bootstrap Equating Errors for the Common-item Nonequivalent Groups Design: A Comparison of Rasch Equating Methods

Mingying Zheng
University of Iowa

Citation: Zheng M. (2022) Bootstrap Equating Errors for the Common-item Nonequivalent Groups Design: A Comparison of Rasch Equating Methods, *British Journal of Education*, Vol.10, Issue 13, pp.56-67

ABSTRACT: *The nonequivalent groups with anchor test (NEAT) equating design are traditionally based on using a single anchor to adjust for differences in test difficulty which is critical to equating test forms in most large-scale testing programs. When tests differ somewhat in content and length, methods based on the item response theory (IRT) model leads to greater stability of equating results. The current study compared standard errors, bias, and root mean square errors using four Rasch IRT equating methods for the nonequivalent groups with anchor test design. The sizes of the equating anchor were employed in all four different Rasch equating methods to investigate how different anchor sizes may impact the test accuracy of the tests by conducting a simulation study.*

KEYWORDS: NEAT design, anchor items, equating accuracy

INTRODUCTION

In most large-scale testing programs, more than one form of a test is often administered at different times and at different locations. Although all such tests should be constructed with the same content and same statistical specifications, both test forms and the samples of test takers cannot be equal, and population parameters will not remain stable over time. The common-item nonequivalent groups design (CINEG) is traditionally based on using anchor items to adjust for differences in test difficulty (Braun & Holland, 1982; Kolen & Brennan, 2014; von Davier, Holland & Thayer, 2004), which is critical to equating test forms in most large-scale testing programs. Substantially and statistically, the anchor test either remains internally or externally in the test forms as internal anchor or external anchor items (Cook & Peterson, 1987). The number of items in a test affects the reliability, and, therefore, possibly the validity of test scores (Messick, 1989). In general, a longer anchor test is considered desirable, because it is more reliable, covers the content better, and it tends to generate fewer random equating errors in a CINEG design (Budescu, 1985). However, when the item parameters of both tests are estimated concurrently, as few as five or six carefully chosen items could perform as satisfactory anchors in IRT equating (Wingersky & Lord, 1984). According to Angoff (1984), a rule of thumb for the minimum number of anchor items is 20 items or 20% of the number, whichever is larger. The presence of anchor items requires additional considerations because these items play a large role in determining the equating function. The properties of these items should be a primary concern when conducting equating studies (Cook & Peterson, 1987). When tests differ somewhat in content and length, methods based on the item response theory (IRT) model lead to

greater stability of equating results (Kolen & Brennan, 2014). According to Lord (1975), IRT methods of equating is useful, especially in increasing the stability of the scales near the extreme values and in reducing drift in equating chains of several test forms. Wang, Qian, and Lee (2013) evaluated the combined effects of reduced equating sample size and shortened anchor test length on item response theory (IRT)-based linking and equating results. The presence of anchor items requires additional considerations because these items play a large role in determining the equating function. The properties of these items should be a primary concern when conducting equating studies (Cook & Peterson, 1987). Kolen and Whiteney (1981) found IRT equating with the one-parameter (Rasch) model to be effective. Although the patterns of standard errors of traditional equating methods (e.g., mean, linear, equipercentile equating methods, etc.) for the single group, random group, and CINEG designs have been studied widely and the results are well-known (Tsai, Hanson, Kolen, & Forsyth, 2001), relatively little information about the standard errors of Rasch IRT equating for the CINEG design is available. One of the few studies conducted by Tsai, Hanson, Kolen, and Forsyth (2001) bootstrap errors of five three-parameter item response theory (IRT) equating methods for the CINEG design were compared. The magnitudes of the estimated standard errors from this study suggested reasonably accurate IRT equating for the CINEG design can be achieved even with an examinee sample size of 500. However, in their study, only standard errors of equating were estimated using different three-parameter IRT equating methods used. Meanwhile in their study, only different sample sizes were compared regarding standard errors of equating without varying the anchor sizes.

The current study extended the research of Tsai, Hanson, Kolen, and Forsyth (2001) by comparing bootstrap equating errors using four Rasch IRT equating methods for the CINEG design: (a) whether Rasch IRT parameters are estimated separately or concurrently, (b) whether a scale transformation is calculated to place parameters for the two forms on a common scale (e.g., Stocking & Lord, 1983), (c) whether the true-score or observed-score equating method is used. The sizes of the equating anchor were examined in all four different methods. Table 1 summarized the characteristics of these four methods in three dimensions using three different anchor test lengths.

Table 1. Four Rasch Equating Methods with Three Different Anchor Test Lengths

Method	*IRT Parameter Estimation*	*Scale Transformation*	*Equating Method*	*Anchor Test Length*
1	Separate Calibration	Yes	True score	16, 20, 24
2	Separate Calibration	Yes	Observed score	16, 20, 24
3	Concurrent Calibration	No	True score	16, 20, 24
4	Concurrent Calibration	No	Observed score	16, 20, 24

Note: Rasch=Rasch Item Response Theory.

METHOD

The Data and Examinees

This study used simulated data with total items of 80. The anchor items sizes of 16, 20, 24, were used based on Angoff's suggestion (1984, p.107). According to Angoff (1984), the rule of thumb for the minimum number of anchor items of 20 anchor items or 20% of the total items in the test. 24 items or 30% of the total items were used to compare the differences. The one-parameter Rasch model is one of the most widely used IRT models (Hambleton, Swamination, & Rogers, 1991). Item characteristic curves for the Rasch model are estimated by the equation

$$p_i(\theta) = \frac{e^{(\theta - b_i)}}{1+e^{(\theta - b_i)}} \ , i = 1, 2, \dots, n \quad (1)$$

where $p_i(\theta)$ is the probability that a randomly chosen examinee with ability θ answering item i correctly, and b_i is the item *i* difficulty parameter.

Simulation Procedure

Methods 1 and 2

In the current study, total number of items was set to 80 with three different anchor sizes of 16, 20, and 24. The values of b_i vary from -2.0 to 2.0 for both Forms X and Y. A sample size of 1,000 examinees with the ability values with mean of 0 and standard deviation of 1 for examinees who took Form X and the ability values with mean of 0 and standard deviation of 0.5 Form Y separately were set. The steps for Methods 1 and 2 for computing bootstrap equating errors of IRT equating for the CINEG design are presented as follows.

1. The true equating functions for three anchor sizes were set to obtain the scales scores for both true and observed score equating methods. Sparse matrices were then obtained for concurrent calibration using *irtoys* package from R (R Core Team, 2017).
2. Bootstrap random samples were drawn for both Form X and Form Y.
3. The package *irtoys* was run separately on each bootstrap random sample to obtain item parameter estimates and θ distributions for Form X and Form Y.
4. The *plink* package from R (R Core Team, 2017) was used to estimate the Stocking and Lord (SL, 1983) scale transformation coefficients θ distributions that were obtained from step 3.
5. The coefficients obtained from step 4 were applied to the entire set of Form X item parameters and ability distributions to produce rescaled Rasch IRT parameter estimates for Form X.
6. The *plink* package from R (R Core Team, 2017) was used to obtain From Y true-score and observed-score equivalents for Form X using estimated item parameter estimates for both forms, with rescaled item parameters on Form X that were obtained from step 5.
7. Steps 1-6 were repeated 1,000 times, and the standard deviation was computed over 1,000 replications to obtain the standard errors of Rasch IRT true-score and observed-score equating at all raw score points for the new form.

In applying the SL (1983) scale transformation method, the common-item parameter estimates were used to obtain equating coefficients A and B. Then, IRT true-score equating (TSE, Method 1) and observed-score equating (OSE, Method 2) were performed with three different anchor test lengths examined. This procedure was replicated 1,000 times to obtain bootstrap estimates of the standard error of equating (SEE).

Methods 3 and 4

A concurrent calibration of both forms was conducted using *irtoys* from R (R Core Team, 2017) to obtain item and ability parameter estimates. Because all of the items and both groups of examinees were put together in the analysis, the parameter estimates were on a common scale. Thus, there was no need to use the SL (1983) scale transformation methods to rescale the IRT parameters on the new form. Finally, IRT true-score (Method 3) and observed-score equating (Method 4) were performed with four different anchor test lengths examined using the package *plink* from R. This procedure was replicated 1,000 times to obtain the bootstrap equating error. The steps for Methods 3 and 4 for computing bootstrap equating errors of Rasch IRT equating for the CINEG design are presented as follows.

1. A bootstrap sample of Form Y examinee item responses and a bootstrap sample of Form X examinee item responses was taken.
2. The package *irtoys* from R (R Core Team, 2013) was run to obtain item parameter estimates for all items and a θ distribution for Form X and Form Y concurrently. In this case, the items that appeared only on Form Y were regarded as not reached for the examinees of Form X, and the items that appeared only on Form X were regarded as not reached for the examinees of Form Y.
3. The *plink* package from R (R Core Team, 2013) was used with estimated item parameters obtained from step 2 to obtain Form Y true-score and observed-score equivalents.
4. Steps 1 to 3 were replicated 500 times, and the standard deviation was computed repeatedly over 1,000 to obtain the standard error of Rasch IRT true-score and observed-score equating methods.

Criteria

At each score point, the standard errors of equating (SEE) were compared directly for four different bootstrap Rasch models (true-score equating with separate calibration and SL scoring transformation; observed-score equating with separate calibration and SL scoring transformation; true-score equating with concurrent calibration; observed-score equating with concurrent calibration). Considering all score points together, the standard error of equating is defined as follows:

$$SEE(x) = \sum_{i-0}^{k}\{se * \left[\widehat{eq_y}\,(x_i)\right]\}, \qquad (2)$$

An RMSE statistic was calculated to take into account both systematic and random errors. Total error of the estimates of standard error can be partitioned into random error (SEE) and systematic error components (bias) as follow:

$$MSE\{se * \left[eq_y\,(x_i)\right]\} = E\{se * \left[\widehat{eq_y}\,(x_i)\right] - \sigma\left[\widehat{eq_y}\,(x_i)\right]\}^2$$

$$= Var\{se * [\widehat{eq_y}(x_i)]\} + \{Bias(se * [\widehat{eq_y}(x_i)])\}^2, \quad (3)$$

where Bias $(se * [\widehat{eq_y}(x_i)]) = E\{se * [\widehat{eq_y}(x_i)] - \sigma[\widehat{eq_y}(x_i)]\}$, E stands for expected value, and σ is the true value of the standard error for a population. The RMSE is the square root of the MSE. That is

$$RMSE\{se * [eq_y(x_i)]\} = \sqrt{MSE\{se * [eq_y(x_i)]\}}$$

$$= \sqrt{Var\{se * [eq_y(x_i)]\} + \{Bias(se * [eq_y(x_i)])\}^2}, \quad (4)$$

Data Analysis

Bootstrap equating errors were computed to examine how four different Rasch IRT equating methods combined with three different anchor test lengths impact the accuracy of the test in a NEAT design. The data analysis procedure for each of the four methods are described as follows.

For Method 1 and Method 2, Rasch IRT parameters for Form X and Form Y are estimated separately and a linear transformation method (Stocking & Lord, 1983) is used to place Rasch IRT parameter estimates for Form X and Form Y scale. Then, Rasch IRT true-score (Method 1) and observed-score (Method 2) equating methods are used to obtain Form Y true-score and observed-score equivalents. For Method 3 and Method 4, Rasch IRT parameters for Form X and Form Y are estimated concurrently. A linear scaling transformation method is not needed because Rasch IRT parameter estimates are already on the same scale. Rasch IRT true-score (Method 3) and observed-score (Method 4) equating methods are used to obtain Form Y true-score and observed-score equivalents. For the first two methods, IRT parameters were estimated using separate calibration and then true-score, observed-score IRT equating with three anchor test lengths employed. For the last two methods, IRT parameters were first estimated using concurrent calibration for Form X and Form Y, then true-score and observed-score equating were performed with three anchor test lengths employed. The number of equating errors associated with these four Rasch equating methods was quantified by the standard errors of equating (SEE), bias, and root mean square errors (RMSE). The bootstrap approach was used to estimate standard errors of IRT equating for each of the four methods.

RESULTS

In the current study, four different Rasch IRT equating methods were used to compare the standard errors of equating with three different lengths of anchor tests used: (a) V_1=20, (b) V_2=16 (20% of the total items), and (c) V_3=24 (30% of the total items). The sample size of the study is 1,000 with 80 total items, the examinees who took either tests of Form X or Form Y.

The Anchor Length of 20

Figure 1 through 3 show the standard errors of equating (SEE), bias, and root mean square error (RMSE) for all four Rasch equating methods with anchor length of 20. The results are summarized as follows:

1. Based on the SEE index, both Rasch IRT true and observed-score equating (separate calibration, Method 1-SEE=0.23, Method 2-SEE=0.24) produced the same smallest standard errors of equating, while both Rasch IRT true and observed-score equating (concurrent calibration, Method 3-SEE=0.33, Method 4-SEE=0.34) produced the same somewhat larger standard errors of equating.
2. Based on the Bias index, the amount of bias varies among four different Rasch IRT equating methods. Among them, Method 1 (separate calibration, Method 1-Bias=0.12) produced the smallest bias of equating followed by Method 2 (separate calibration, Method 2-Bias=0.18) while Method 3 (concurrent calibration, Method 3-Bias=0.32, Method 4-Bias=0.30) produced the largest bias of equating.
3. Based on the RMSE index, both Rasch IRT true and observed-score equating (separate calibration, Method 1-RMSE=0.28, Method 2-RMSE=0.29) produced the same smallest standard errors of equating, while both Rasch IRT true and observed-score equating (concurrent calibration, Method 3-RMSE=1.12, Method 4-RMSE=1.08) produced the same somewhat larger root mean square errors (RMSE).
4. Both Rasch true and observed score equating methods with separate calibration produced same smallest SEE, bias, and RMSE.
5. The SEE was normally distributed for all four Rasch Equating methods. The magnitude of the bias of equating were quite different across four methods. Method 3 produced the greatest bias.

The RMSE round the equated scores of 42 was approximately 0.25. Thus, the equating results around the score of 42 appear to be reasonably accurate with anchor length of 20.

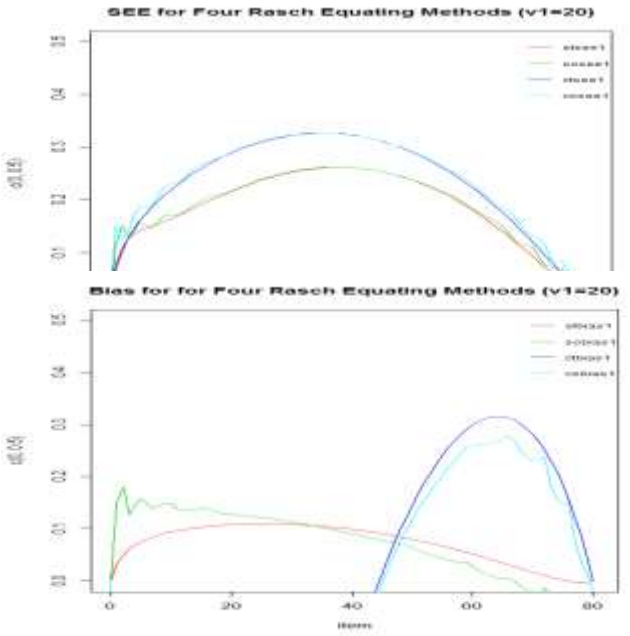

Figure 1. Standard Errors of Equating for the Four Rasch Equating Methods (V1=20). Stsee2=separate true-score standard errors of equating; sosee1= separate observed-score standard errors of equating; ctsee1= concurrent true-score standard errors of equating; cosee1= concurrent observed-score standard errors of equating.

Figure 2. Bias of Equating for the Four Rasch Equating Methods (V1=20). stbias1=separate true-score bias of equating; sobias1= separate observed-score bias of equating; ctbias1= concurrent true-score standard errors of equating; cobias1= concurrent observed-score standard errors of equating.

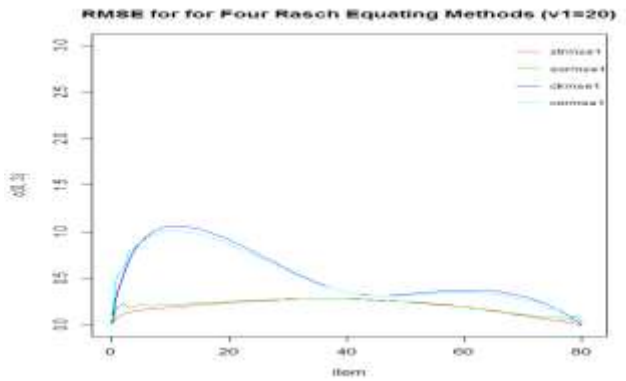

Figure 3. RMSE of Equating for the Four Rasch Equating Methods (V1=20). strmse1=separate true-score root mean square error (rmse) of equating; sormse1= separate observed-score rmse of equating; ctrmse1= concurrent true-score rmse of equating; cormse1= concurrent observed-score rmse of equating.

The Anchor Length of 16

Figure 4 through 6 show the standard errors of equating (SEE), bias, and root mean square error (RMSE) for all four Rasch equating methods with anchor length of 16. The results are summarized as follows:

1. Based on the SEE index, both Rasch IRT true and observed-score equating (separate calibration, Method 1-SEE=0.43, Method 2-SEE=0.44) produced the same somewhat larger standard errors of equating, while both Rasch IRT true and observed-score equating (concurrent calibration, Method 3-SEE=0.32, Method 4-SEE=0.33) produced the same somewhat smallest standard errors of equating.
2. Based on the Bias index, the amount of bias varies among four different Rasch IRT equating methods. Among them, Method 1 (separate calibration, Method 1-Bias=0.07) produced the smallest bias of equating followed by Method 2 (separate calibration, Method 3-Bias=0.17) while Method 3 (concurrent calibration, Method 3-Bias=0.42; Method 4-Bias=0.40) produced the largest bias of equating.
3. Based on the RMSE index, both Rasch IRT true and observed-score equating (separate calibration, Method 1-RMSE=0.41, Method 2-RMSE=0.42) produced the same smallest root mean square errors (RMSE) of equating, while both Rasch IRT true and observed-score equating (concurrent calibration, Method 3-RMSE=0.91, Method 4-RMSE=0.89) produced the same somewhat larger root mean square errors (RMSE).
4. Both Rasch true and observed score equating methods with separate calibration produced same somewhat bias, and RMSE.
5. The SEE was normally distributed for all four Rasch Equating methods. The magnitude of the bias of equating were quite different across four methods. Method 3 produced the greatest bias.
6. The RMSE around the equated score of 42 was approximately 0.25. Thus, the equating results around the score of 42 appear to be reasonably accurate with anchor length of 16.

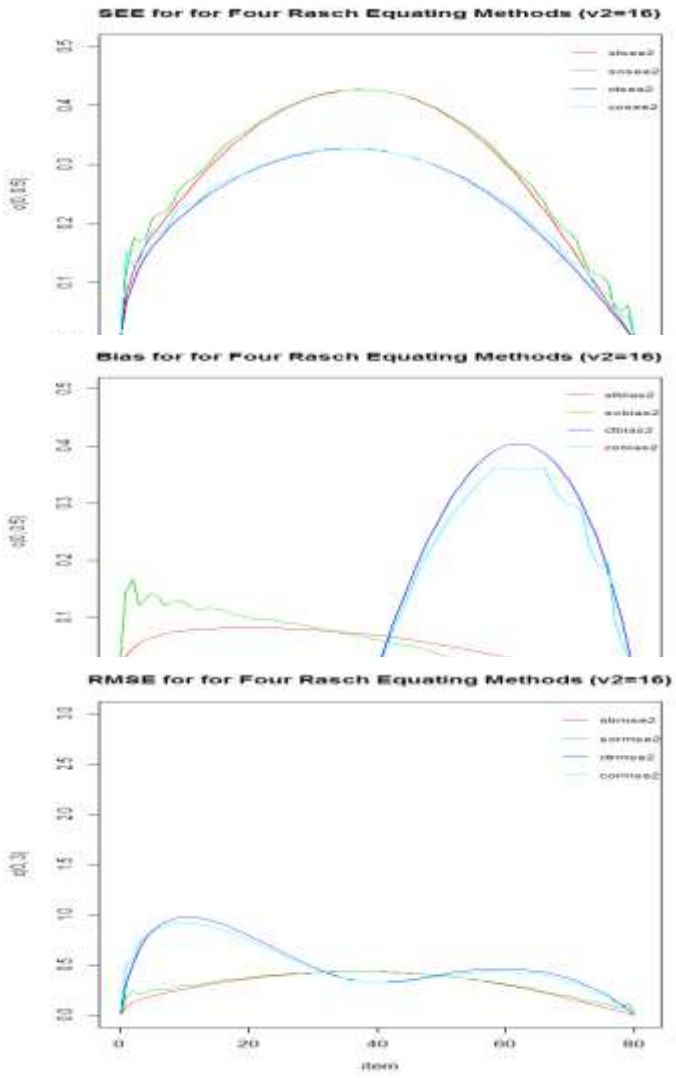

Figure 4. Standard Errors of Equating for the Four Rasch Equating Methods (V2=16). stsee2=separate true-score standard errors of equating; sosee2= separate observed-score standard errors of equating; ctsee2= concurrent true-score standard errors of equating; cosee2= concurrent observed-score standard errors of equating.

Figure 5. Bias of Equating for the Four Rasch Equating Methods (V2=16). stbias2=separate true-score bias of equating; sobias2= separate observed-score bias of equating; ctbias2= concurrent true-score standard errors of equating; cobias2= concurrent observed-score standard errors of equating.

Figure 6. RMSE of Equating for the Four Rasch Equating Methods (V2=16). Strmse2=separate true-score root mean square error (rmse) of equating; sormse2= separate observed-score rmse of equating; ctrmse2= concurrent true-score rmse of equating; cormse2= concurrent observed-score rmse of equating.

The Anchor Length of 24

Figure 7 through 9 show the standard errors of equating (SEE), bias, and root mean square error (RMSE) for all four Rasch equating methods with anchor length of 24. The results are summarized as follows:

1. Based on the SEE index, both Rasch IRT true and observed-score equating (separate calibration, Method 1-SEE=0.25, Method 2-SEE=0.26) produced the same somewhat smallest standard errors of equating, while both Rasch IRT true and observed-score equating (concurrent calibration, Method 3-SEE=0.95, Method 4-SEE=0.96) produced the same somewhat larger standard errors of equating.

2. Based on the Bias index, the amount of bias varies among four different Rasch IRT equating methods. Among them, Method 1 (separate calibration, Method 1-Bias=0.05) produced the smallest bias of equating.
3. Based on the RMSE index, both Rasch IRT true and observed-score equating (separate calibration, Method 1-RMSE=0.61, Method 2-RMSE=0.62) produced the same smallest standard errors of equating, while both Rasch IRT true and observed-score equating (concurrent calibration, Method 3-RMSE=1.46, Method 4-RMSE=1.44) produced the same somewhat larger root mean square errors (RMSE).
4. Both Rasch true and observed-score equating methods with separate calibration produced same smaller bias, and RMSE.
5. The SEE was normally distributed for all four Rasch Equating methods. The magnitude of the bias of equating were quite different across four methods.

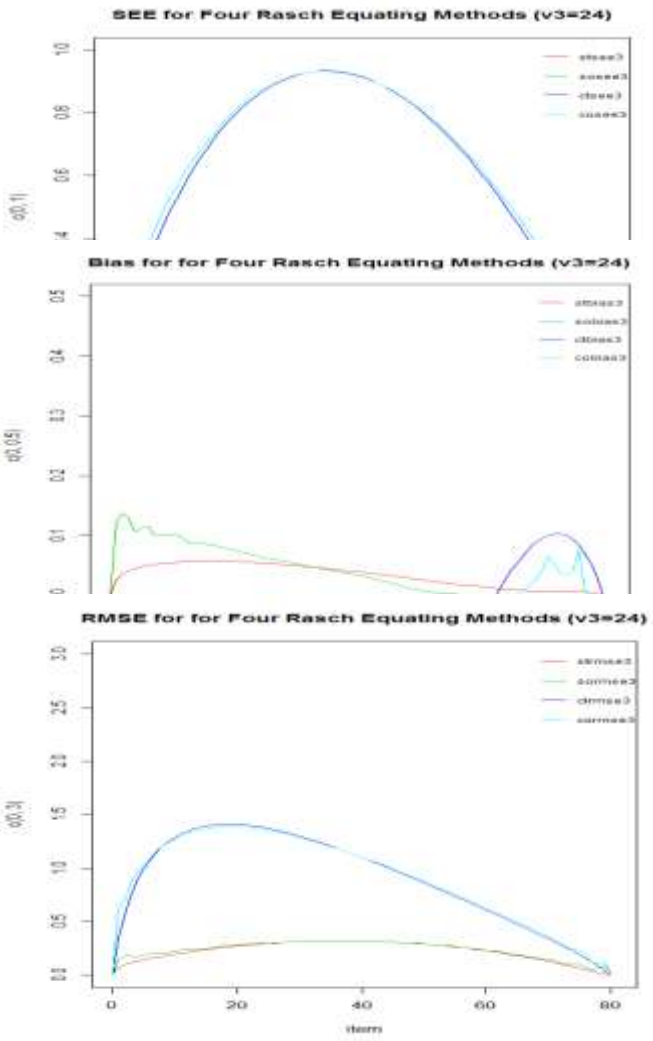

Figure 7. Standard Errors of Equating for the Four Rasch Equating Methods (V3=24). stsee3=separate true-score standard errors of equating; sosee3= separate observed-score standard errors of equating; ctsee3= concurrent true-score standard errors of equating; cosee3= concurrent observed-score standard errors of equating.

Figure 8. Bias of Equating for the Four Rasch Equating Methods (V3=24). stbias3=separate true-score bias of equating; sobias3= separate observed-score bias of equating; ctbias3= concurrent true-score standard errors of equating; cobias3= concurrent observed-score standard errors of equating.

Figure 9. RMSE of Equating for the Four Rasch Equating Methods (V3=24). strmse3=separate true-score root mean square error (rmse) of equating; sormse3= separate observed-score rmse of equating; ctrmse3= concurrent true-score rmse of equating; cormse3= concurrent observed-score rmse of equating.

DISCUSSION

In this study, bootstrap procedures were described for estimating standard errors of four different Rasch IRT equating methods in three dimensions. Three different anchor test lengths were employed in all four methods. Similar procedures can be used for estimating standard errors of IRT equating in other simulation testing studies. The estimated standard errors can be used to plan the anchor test sizes required for IRT equating and to document the amount of error in IRT equating results.

The magnitudes of the estimated standard errors, bias, and RMSE of equating found in this study suggested that when IRT parameter estimation was conducted separately by form, the standard errors of IRT true and observed score equating were similar, which confirmed the results from Tsai, Hanson, Kolen, and Forsyth's study (2001). Meanwhile, the amount of SEE is quite small when Rasch IRT parameter estimation was conducted separately. The magnitudes of the estimated standard errors found in this study suggested that reasonably accurate IRT equating for the CINEG design can be achieved with an anchor size of 20 or 20% of the total items. However, increasing anchor test sizes does not contribute further to lowering the SEE and RMSE significantly, indicating the length of the anchor tests is not the most critical factor to improve testing accuracy. More investigation can be focused on the quality of the anchor items.

The methods based on the concurrent calibration of Rasch IRT parameters had more SEE than did the methods based on separate calibration, which contradicted the results from Tsai, Hanson, Kolen, and Forsyth's study (2001).

This study compared the SEE, bias, and RMSE using four Rasch IRT equating methods with three different anchors employed in all four methods. As the total equating error (RMSE) consists of random error (SEE) and systematic error (bias), the relative performance of the methods studied in this article with regard to random error would have changed to some extent when total equating errors were considered.

In sum, the IRT equating methods used in the current study can be used as the reference when different calibrations are needed in large-scale tests. The results from the study regarding the anchor test length can also be recommended in testing systems.

Limitations and Future Study

This study compared four Rasch IRT equating methods in three dimensions when three different anchor test lengths employed for CINEG design. The random error, systematic error, and total error of equating were estimated. The entire study used the simulated data instead of empirical data. The results cannot be generalized to empirical studies. Future investigations should be made using empirical data to validate the findings from this study. In this study, only one-parameter (Rasch) IRT models were used. Further studies can be conducted using different IRT models. In the current study, only anchor test length is investigated, different factors that may affect the accuracy of test forms should be investigated such as scale stability, anchor item quality, and total test length, etc. should

also be examined. In addition, the traditional equating methods and IRT equating methods can be used to compare random, systematic error, and total errors to investigate the testing accuracy.

References

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

Braun, H. I., & Holland, P. W. (1982). Observed score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), Test equating (pp. 9–49). New York, NY: Academic Press.

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. *Journal of Educational Measurement, 22,* 13–20.

Cook, L. L., & Petersen, 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.

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking*. New York, NY: Springer New York. Retrieved from http://link.springer.com/10.1007/978-1-4939-0317-7

Kolen. M. J. & Whiteney, D. R. (1981). Comparison of four procedures for equating the tests of general educational development. *Journal of Educational Measurement, 19,* 279-294.

Lord, F. M. (1975). A survey of equating methods based on item characteristic curveTheory. *Research Bulletin,* 75-13. Princeton, NJ: Educational Testing Service.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale,NJ: Lawrence Erlbaum.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, March* 5-11.

Moses, T., Deng, W., & Zhang, Y.-L. (2010). The use of two anchors in nonequivalent groups with anchor test (NEAT) equating. *ETS Research Report Series*, *2010*(2), i–33.

Moses, T., & Kim, S. (2007). Reliability and the Nonequivalent Groups with Anchor Test Design. *ETS Research Report Series*, *2007*(1), i–40.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ricker, K. L., & Davier, A. A. (2007). The impact of anchor test length on equating results in a nonequivalent groups design. *ETS Research Report Series*, *2007*(2), i–19.

Sinharay, S., & Holland, P. W. (2009). The missing data assumptions of the nonequivalent groups with anchor test (NEAT) design and their implications for test equating. *ETS Research Report Series*, *2009*(1), i–53.

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

Tsai, T.-H., Hanson, B. A., Kolen, M. J., & Forsyth, R. A. (2001). A Comparison of Bootstrap Standard Errors of IRT Equating Methods for the Common-Item Nonequivalent Groups Design. Applied Measurement in Education, 14(1), 17–30. http://doi.org/10.1207/S15324818AME1401_03

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

Wang, L., Qian, J., & Lee, Y.-H. (2013). Exploring Alternative Test Form Linking Designs With Modified Equating Sample Size and Anchor Test Length. *ETS Research Report Series, 2013*(1), i–17.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8,* 347-364.