KURAM VE UYGULAMADA EĞİTİM BİLİMLERİ EDUCATIONAL SCIENCES: THEORY & PRACTICE

Received: December 8, 2014
Revision received: December 29, 2015
Accepted: January 4, 2016
OnlineFirst: March 30, 2016

www.estp.com.tr
DOI 10.12738/estp.2016.2.2762 ♦ April 2016 ♦ 16(2) ♦ 647-668

Research Article

Investigating Test Equating Methods in Small Samples Through Various Factors*

Semih Aşiret[1]
Mersin University

Seçil Ömür Sünbül[2]
Mersin University

Abstract

In this study, equating methods for random group design using small samples through factors such as sample size, difference in difficulty between forms, and guessing parameter was aimed for comparison. Moreover, which method gives better results under which conditions was also investigated. In this study, 5,000 dichotomous simulated data consistent with the three-parameter logistic model were produced for each form (X and Y) by manipulating the factors' levels for random group design. In order to equate two test forms, the random groups design was used in this study. Simulated test forms were equated using the equating methods of identity, mean, linear, circle-arc, and 2- and 3-moments pre-smoothed equipercentile for different sample sizes (10, 25, 50, 75, 100, 150, 200) through 100 replications. The results obtained from this simulation study were evaluated based on the criterion of equating error (RMSE). The findings indicated that in the case where the sample size was 50 or more and the difficulty difference between forms was 0.4, equating the forms was concluded to give better results than not equating. Moreover, the circle-arc and mean-equating methods produced lower equating errors than other equating methods for small-sample equating under most of the conditions studied.

Keywords

Test equating • Equating error • Small samples

* This study was developed from Semih Aşiret's master's thesis that was completed under the advice of Assistant Professor Seçil Ömür Sünbül at Mersin University, Institute of Educational Sciences.

1 **Correspondence to:** Semih Aşiret, Department of Measurement and Evaluation in Education, Faculty of Education, Mersin University, Yenişehir Campus, Mersin 33169 Turkey. Email: semihasiret@gmail.com

2 Department of Measurement and Evaluation in Education, Mersin University, Mersin 33169 Turkey. Email: secilomur@gmail.com

Citation: Aşiret, S., & Sünbül, S. Ö. (2016). Investigating test equating methods in small samples through various factors. *Educational Sciences: Theory & Practice, 16*, 647-668.

In education and psychology, tests and scales are widely used for monitoring the learning levels of examinees, placing them within upper-instructional levels, selecting staff members, conducting guidance, and performing clinical services. Important decisions are made about examinees by taking into account their test or scale scores. However, the scores obtained from tests and scales have to be accurate in order to make a fair decision about examinees.

Sometimes tests and scales are administrated at different times; due to security reasons, different parallel test forms can be used at the same time to overcome this problem. However, this situation causes some problems. Even if test developers construct test forms whose content and statistical characteristics are the same, the forms can have different difficulty levels. While some test forms consist of easy items, others may have difficult items that can cause examinees' scores to differentiate. To overcome this problem, test forms that are constructed at different times should be arranged as different forms of the same test. This case, however, leads to concerns about the simplicity or difficulty of the test forms. The scores of the examinees who are tested at different times cannot be directly compared. When two different forms measuring the same construct are administrated to different groups of examinees, the difficulty of the items on the test forms may not be equal. Test equating is used to overcome these difficulties and can interchangeably interpret the scores obtained from test forms (Kolen & Brennan, 2004; vonDavier, Holland, & Thayer, 2004).

Kolen and Brennan (2004) defined test equating as the statistical process used for adjusting scores obtained from test forms so that these scores can be used interchangeably. Crocker and Algina (1986) have defined test equating as a process that establishes equivalent scores from two different measurement instruments; they pointed out that when the percentiles corresponding with the X and Y scores obtained from different tests that have equal reliability and measure the same construct are equal, the tests that these X and Y scores were obtained from are equal. Angoff (1984) defined test equating as the process of converting the system of units from one test form to the system of units of another test form, pointing out that scores obtained from different forms are equated after the scores are transformed. Consequently, test equating emerged due to the fact that two or more tests forms which measure the same content and construct can produce different scores for the same examinees.

Certain requirements must be satisfied to equate two test forms. There are many different views related to these requirements in the literature. Hambleton and Swaminathan (1985) listed these requirements as the properties of *symmetry*, *same specifications*, *equity*, and *group*. With symmetry, when transforming the scores from Form X to Form Y, the inverse of this transformation process should also be valid (Kolen & Brennan, 2004). According to the property of same specifications, the test forms to be equated are required to have the same content and statistical properties. The scores obtained from an equation which ignores these statistical properties cannot be used interchangeably (Kolen & Brennan, 2004). In the property of equity of equating, as proposed by Lord (1980), indifference towards whether Form X or Form Y is administered to the examinees must be claimed. However, this property holds for when test forms are identical. When identical forms are constructed, it is not necessary to equate forms (Crocker & Algina, 1986; Kolen & Brennan, 2004). When claiming the property of group invariance, equating test forms will be independent of examinee group; it does not matter which group is chosen for calculating the equating function between the scores from Form X and Form Y (Kolen & Brennan, 2004; Öztürk & Anıl, 2012).

It is necessary that either common examinees are administrated the two tests or that common items are placed in the two tests to collect data in test equating; for this reason, data collection design has been developed. Data collection design is a plan for collecting the data needed for equating. The data collection design may be categorized as a common-examinees design or as a common-items design. The single-group design, random-groups design, and counterbalanced single-group design are common designs used by examinees (Kolen & Brennan, 2004). In this study, the random-groups design was used.

In the random-groups design, it is necessary to equate the test forms that have been administered to similar examinees (Hambleton & Swaminathan, 1985). In practice, however, the test forms are randomly assigned to equivalent groups. Two samples are randomly and independently drawn from a common population of examinees; Form X is administrated to the first sample, and Form Y is administrated to the second sample (vonDavier et al., 2004). Examinees are assigned the forms randomly. In this design, examinees receive one form, as opposed to the single-group design. Thus, when compared with other designs, the random-groups design enables time to be saved, and it can be applied more practically. Furthermore, multiple forms can be administrated simultaneously (such as Form C, Form D). In the random-groups design, the difference between the group performances of examinees who have taken different forms reveals the difference in difficulty between the test forms (Kolen & Brennan, 2004).

Kolen and Brennan (2004) stated that to determine how to equate a design, one should consider practical situations, such as the complexity of administering the test, challenges in test development, and the ability to meet statistical assumptions. Crocker and Algina (1986) pointed out that practicability was the main criteria for determining how to equate a design. As common items are not required to represent the entire content of the test, constructing test forms in the random-groups design is less complex than other equating designs. The random-groups design has the least problems with statistical assumptions owing to the random assignment of forms to examinees, and there is no problem with the effect of order. Consequently, the random-group design was preferred in this study because of the ease of developing and administering the test, and because it has the least problems with required statistical assumptions.

Equating is a statistical process which is used to transform scores from one test form to the scale of another test form. There are many methods related to transforming forms (Dorans, Moses, & Eignor, 2000). Equating methods have been classified according to a theory based on methods of classical test theory and item-response theory (IRT). In this study, methods based on classical test theory have been used; as such, these methods will hence be discussed briefly.

Identity Equating

This method may not be considered as an equating method because the scores between forms are not transformed. This means they are considered already equal and do not need equating. The mathematical function of the identity-equating method is formulized in Equation 1.

$$y = \mathrm{IDy}(x) = x \quad (1)$$

In Equation 1, x refers to the raw score obtained from Form X, and y is the raw-score equivalent of x for Form Y.

Skaggs (2005) and Kim, vonDavier, and Haberman (2006) stated that the random error would be zero because the scores obtained from the forms had a one-to-one equivalence. On the other hand, when forms are not parallel, systematic errors (such as bias) increase. Kolen and Brennan (2004) suggested that when test forms are considered parallel, the identity-equating method should be preferred.

Mean Equating Method

This considers that Form X is differentiated from Form Y in difficulty by a constant amount over the score scale (Kolen & Brennan, 2004). There is no difference in the ability levels of examinees who take the forms. In the mean-equating method, the scores of the two forms are determined using Equations 2 and 3.

$$x\text{-}\mu_x = y - \mu_y \quad (2)$$

$$m_y(x) = y = x - \mu_x + \mu_y \quad (3)$$

In these equations, x is the score from Form X, is the mean of Form X, y is score from Form Y, is the mean of Form Y, and is the score transformed from x on Form X to Form Y by using mean equating.

Linear Equating

Crocker and Algina (1986) pointed out that the linear-equating method was based on the assumption that the distributions of scores on Form X and Form Y were the same, but their means and standard deviations were different. Linear equating is used when the standard scores derived from these forms are considered to be equivalent. Donlon (1984) stated that if groups of examinees who had taken different forms of a test had equal ability levels, linear equating could be implemented.

Angoff (1984, p. 564) defined linear equating as scores being equivalent when the scores on two test forms correspond to the same standard-score deviations. Angoff formulized this situation in Equation 4:

$$\frac{(y-\mu_y)}{\sigma y} = \frac{(x-x_\mu)}{\sigma x} \tag{4}$$

Equation 5 is derived by rearranging the fourth equation.

$$Liny(\mathrm{x}) = \frac{\sigma_y}{\sigma_x} + \mu_y \frac{\sigma_y}{\sigma_x} \mu_x \tag{5}$$

Equation 5 can be expressed as Y = (AX) + B, where refers to slope of the line and B = refers to the intercept point of the line. This score transformation is symmetric in contrast to regression equating (Angoff, 1984).

Equipercentile Equating

This method is recommended when the score distributions of the forms are different. In equipercentile equating, Form X may be more difficult than Form Y for high and low scores, however it may be less difficult for middle scores (Kolen & Brennan, 2004). Examinees' scores on forms X and Y are equated with respect to their corresponding percentile ranks (Kolen, 1988). If the distribution of scores on Form X which transformed to scores on Form Y is equal to the distribution of scores on Form Y, the equating function between the two forms is called the equipercentile-equating function (Kolen & Brennan, 2004).

In equipercentile equating, percentile ranks for each form are first calculated. Scores that correspond to the same percentile rank are equivalent **(**Kolen, 1988; Kolen & Brennan, 2004; Livingston, 2004). However, these equating processes are based on the assumption that the test scores are continuous variables. Actually, test scores are discrete variables. When test scores are discrete variables, the equipercentile-equating function cannot be used. To overcome this limitation, discrete variables are viewed as continuous variables by transforming scores into percentiles or percentile groups (Kolen & Brennan, 2004). In equipercentile equating, one issue can arise where an examinee receives no score. To handle this situation, the middle point of the range of scores that correspond to the same percentile group is chosen as being equivalent.

Smoothing in Equipercentile Equating Method

In the equipercentile-equating method, the examinees who will take the test forms are sampled from one or more populations. While drawing these samples, some irregularities can appear as a result of sampling errors when the raw score distribution is graphed (Kolen & Brennan, 2004). Sampling errors can be minimized by increasing the sample size. Kolen and Brennan (1995) suggested that a sample size of 1,500 is ample for the equipercentile-equating method. However, it may not be possible to attain this sample size all the time. For this reason, smoothing methods are used (Cui & Kolen, 2009; Donlon, 1984). Smoothing is used for minimizing sampling errors. The process that produces a new observed-score distribution by eliminating irregularities without changing the distribution's range, shape, or location is called smoothing (Livingston, 2004).

Smoothing is divided into two parts (pre-smoothing and post-smoothing) in accordance with when it happens. The process which smoothes the raw scores' frequency distribution before applying the equipercentile-equating method is called pre-smoothing (Kolen & Brennan, 2004; Livingston, 2004). The rolling average method, log-linear model, and strong true-score theory are types of pre-smoothing methods. When the smoothing of the score distributions is applied to transformations obtained after equipercentile equating has been performed, it is called post-smoothing (Kolen & Brennan, 2004; Öztürk, 2010). The cubic-spline method and polynomials can be used as post-smoothing methods. In this study, the polynomial log-linear pre-smoothing model has been used. Equation 6 is used for the polynomial log-linear model.

$$\log[N_x\ f(x)]=\omega_0+\omega_1 x + \omega_2 x^2+ \quad +\omega_c x^c \tag{6}$$

In this equation, N refers to sample size, f(x) refers to the model as applied to the relative frequency of distribution, C refers to the degree of the lower-order polynomial, and refers to the estimated parameters of the polynomial function.

Circle-Arc Method

Livingston and Kim (2009b) suggested a new method for equating test forms in small samples. They clarified that this method could equate by reducing the number of parameters for estimating the equating relationship in a small sample by neglecting some assumptions. Thus they developed the circle-arc equating method. This method is a strong model which does not assume the equating relationship to be linear. Livingston and Kim (2010) divided the circle-arc method into two categories, the symmetric circle-arc method and the simplified circle-arc method. In the symmetric circle-arc method, two pre-specified endpoints and a middle point that has been determined empirically are fitted into the arc of a circle. In the simplified circle-arc method, equating transformation is decomposed into linear and curvilinear components. In the circle-arc method, the lower endpoint is the lowest meaningful test score in each form. The upper endpoint is the maximum possible score for the test form. The middle point on the curve is the point in the middle of the distribution of scores. The mean of the scores from the test forms is determined as the middle point for the scores obtained from the single-group design, balanced-groups design, and equivalent-groups design. If these three points are on a straight line, this line is called the estimated equating curve. On the other hand, if these three points do not fall on a straight line, they determine a *circle-arc* (Livingston & Kim, 2009b). The equating function has linear and curvilinear components. Both components are calculated separately. The sum of the linear and curvilinear components composes the circle-arc equating function (Livingston & Kim, 2009b).

Equating Error

Being an equating test form, one important point to consider is statistical errors. Kolen and Brennan (2004) divided equating errors into two sources, random error and systematic error. A random equating error (sampling error) occurs when the parameters of a sample that are drawn from the whole population, such as the mean, standard deviation, and percentile rank, are estimated (Kolen, 1988). Random errors may also be defined as the difference between the estimated equating relationship for the samples and for the whole population. Kolen (1988) pointed out that random sample errors can be minimized by increasing the sample size and selecting an appropriate equating design. When the whole population is available, no random errors will be present (Kolen & Brennan, 2004). Systematic errors occur when there are violations of the statistical assumptions or conditions of the equating methods. For example, failing due to being fatigued or getting a high score due to practicing causes systematic errors in the single-group design. Aside from this, if the spiraling process in a single-group design is unable to group comparably, systematic errors result. As a result, if Form X and Form Y differ in difficulty, content, and reliability, systematic errors can be concluded to appear (Kolen, 1988).

Random equating errors can be reduced by increasing the sample size. However, increasing the sample size does not make systematic errors decrease. If smoothing procedures are not applied accurately, equating errors are produced. When smoothing is applied accurately, random errors are reduced; on the other hand, when smoothing is overly applied, systematic errors are introduced (Kolen & Brennan, 2004). Furthermore, random errors and systematic errors can be increased by increasing the number of forms to be equated.

Criterion Equating

Criterion equating is required in order to determine if equating has been done accurately or not. In this study, data for the whole population is available because it uses simulated data. For this reason, the large sample criteria method has been selected for criterion equating in this study. In the large sample criteria method, a large sample representing the population is selected. After this, smaller samples are drawn from the same population and the results are compared with the results from the large sample (Kolen & Brennan, 2004).

Purpose and Significance of the Research

In this study, comparing equating methods for the random-group design was intended using small samples through factors such as sample size, average difference in difficulty between forms, and guessing parameters. When using different forms of a test which are composed of items that measure the same constructs, these forms should initially be interchangeably equated using the scores obtained from the forms. Many test-equating studies in the literature have been typically based on large samples, and their equating methods were suggested for large samples. The larger a sample size is, the better the sample represents the population and the greater the accuracy of equating. However, large sample sizes may not always be available. Hence, test forms should be equated when samples are unavoidably small. Parshall, Houghton, and Kromrey (1995) claimed as well that test equating can be required even with a small sample size. For example, teacher-made tests or exams for small samples are administrated to students in universities or in course centers. If the sample size is substantially small, some problems can occur in equating, such as being unrepresentative of the population or violating the assumptions of equating methods. Livingston and Kim (2009b) suggested using strong models that reduce the number of parameters for estimating from the data to deal with the problem of small samples. In the case of a small sample, one must decide whether the forms will be equated or not. If equating is required, one must decide what conditions are required and which equating methods should be preferred.

There have been few research studies in the literature on equating small samples (Hanson, Zeng, & Colton, 1994; Heh, 2007; Kim et al., 2006; Livingston, 1993; Parshall et al., 1995; Skaggs, 2005). In these studies, pre-smoothed equipercentile equating methods (identity, mean, linear, and log-linear with different moment values from 2 to 6) were typically used. However, the circle-arc equating method, developed by Livingston and Kim (2009b), has only been used in recent years to equate test forms in small-sample equating. Livingston and Kim (2009b) suggested that researchers should test the circle-arc equating method and examine its effectiveness and accuracy. Therefore, the circle-arc method along with other methods was included in this study. Kim et al. (2006) pointed out that the major decision in small samples is whether or not to use the identity-equating method. This decision depends on many factors such as sample size, equating design, test length, guessing parameters, and difference of difficulty between forms. When investigating the relevant literature, one important factor that has clearly had an effect on equating is sample size. Livingston (1993), Hanson et al. (1994), Skaggs (2005), Heh

(2007), Livingston and Kim (2009b), and Devdass (2011) mainly used sample sizes that ranged from 25 to 300 in their studies related to small-sample equating. Babcock, Albano, and Raymond (2012) studied with a small sample of 20; Parshall et al. (1995) studied with small sample of 15; Kim, vonDavier, and Haberman (2008) and Livingston and Kim (2010) studied with small samples of 10 in their studies on small-sample equating. As a result of Livingston and Kim's (2010) study, the circle-arc equating method was stated to have the lowest equating error with sample sizes of 25 and smaller. For this reason, the sample sizes in this study were manipulated to range from 10 to 200. When test forms are nearly parallel, the difference of difficulty between forms is too small. In practice, however, constructing nearly parallel test forms is difficult. Particularly in the case of item-parameter estimations for small samples, constructing parallel test forms is certainly much more difficult. In the literature, it has not been clear what should be done when sample size is small and the difference in form difficulty is large. Differences in test form difficulty have a substantial effect on the estimation errors in equating (Heh, 2007). Skaggs (2005) stated that the degree of average difference in form difficulty affects which equating method should be selected. In the literature, the unit of average difference in form difficulty is often referred to as a standardized mean difference (SMD). When examining the effect of average difference in difficulty between forms in order to equate them, Parshall et al.'s (1995) average difference ranged from 0.0 to 0.4 SMD; Kolen and Brennan's (2004) ranged from 0.1 to 0.6 SMD; Heh's (2007) ranged substantially from 0.0 to 0.75 SMD; Livingston and Kim's (2010) ranged from 0.17 to 0.30 SMD; Devdass's (2011) ranged from 0.10 to 0.25 SMD; and Babcock et al.'s (2012) ranged from 0.0 to 0.75 SMD. In this study, test forms differed in difficulty from 0.1 to 0.4 SMD and from 0.1 to 0.7 SMD in order to examine the effect of mean difference in form difficulty on equating. When reviewing the studies in the literature on the effect of guessing parameters on equating, Bozdağ (2010) pointed out that equating methods differed in equating scores with the probability of success and equating scores probability of failure. However, in the literature, there have been no studies related to probability of success in small samples. In this study, guessing parameters differed between 0.0 and 0.25 in order to examine its effect on equating. The current study eventually aimed to determine which equating methods have the lowest equating error across different levels of sample size, mean difference between form difficulty, and guessing parameters. Therefore, this study is expected to be able to give practical guidance to people who study test equating in small samples. Another expectation is that it will contribute to the field as it is the first research related to test equating in small samples in Turkey. Connected with the specific purpose of this study, the following research questions were addressed:

- What are the main effects of sample size, mean difference between form difficulty, and guessing parameters on the equating errors (RMSE) of the various equating methods?

- What are the interaction effects of sample size, mean difference between form difficulty, and guessing parameters on the equating errors (RMSE) of the various equating methods?

Method

The purpose of this study was to compare the test equating methods under various factors such as sample size, average difference in difficulty between forms, and guessing parameters. Hence, it is expected to contribute to theoretical studies related to test equating in small samples by estimating the equating error (RMSE) of various equating methods under these factors using the random group design. This study is a theoretical research in this respect.

Data Collection

In this study, simulated data were generated by the three-parameter logistic model (3PLM) based on item response theory (IRT) with respect to manipulated factors and their levels. Data were generated using the R 3.1 software program. Firstly, dichotomous data for a population of 5,000 individuals were generated under the random groups design for each form (X and Y). Each form consisted of 30 dichotomous items. Ability distribution of examinees was obtained through normal distribution with a mean of 1 and standard deviation of 0; item discrimination parameters were obtained through normal distribution with a mean of 1.00 and standard deviation of 0.05.

The *b* parameters (difficulty difference) were obtained by uniform distribution where the mean was 0 and standard deviation was 1 to generate the data for Form X. By manipulating the *c* parameter in four stages (0, 0.1, 0.2, and 0.25), four different types of Form X were generated in total. To generate the data for Form Y, *b* parameters for Form Y were obtained by adding 0.1, 0.4 and 0.7 to the *b* parameters of Form X. Moreover, the *c* parameter of Form Y was also differentiated in four stages (0, 0.1, 0.2, 0.25). Consequently, 12 different types of Form Y were generated in total.

Data Analysis

Twenty forms consistent with the observed factors and their levels were simulated overall. Dichotomous data in each row were added to each form to obtain the raw scores. In this study, the random-groups design was used. Form X was considered to be the original form, and Form Y was considered to be the new form. Form X was transformed to Form Y.

The equate package for R 3.1 was used to equate the forms. The equipercentile equating method was used as the equating criterion to control accuracy of equating over the whole population. One hundred different random samples were drawn for each sample size. Form X was equated to the Form Y scale for each replication, and equated scores were computed. After the equating process, equating error (RMSE) was calculated for each method for all sample sizes. The factors and factor levels in this study are shown in Table 1.

The main and interactive effects of the factors studied (sample size, mean difference in form difficulty, guessing parameters) on equating error were examined and graphed.

The equating error, or root-mean square error (RMSE), is the square root of the total equating error variance over 100 replications. The RMSE is equal to the square-root of the sum of the square of standard error equating and the square of equating bias (Kim et al., 2006; Skaggs, 2005). The function for equating error has been given in Equation 7.

$$RMSE_i = \sqrt{(\hat{d}_i^2 + \hat{sd}_i^2)} \tag{7}$$

To determine accuracy of equating methods, the RMSE values for equating and the equating error for the identity-equating method were compared.

Table 1
Factors and their Levels in the Study

Factor	Factor Levels	Values
Sample Sizes	7	10 (N1), 25 (N2), 50 (N3), 75 (N4), 100 (N5), 150 (N6), 200 (N7)
Mean Difference in Difficulty Between Forms	3	0.1, 0.4, 0.7
Parameter Estimate (c)	4	0.00, 0.10, 0.20, 0.25
Equating Methods	6	Identity Equating (ID), Mean Equating (ME) , Linear Equating (LIN), 2-Moments Log-Linear Pre-smoothing Equipercentile Equating (LLC2), 3-Moments Log-Linear Pre-smoothing Equipercentile Equating (LLC3), Circle-Arc Equating (C)

Findings

The Main Effect of Sample Size on Equating Error

The RMSE values for equating methods across different sample sizes are shown graphically in Figure 1.

Investigating Figure 1 shows that as sample size increased, the RMSE values of the different equating methods decreased, with the exception of the identity method. It is clear in Figure 1 that the identity method had the least amount of error with a sample size of 10. This figure shows that some of the equating methods were feasible with a sample size of 25: The circle-arc and mean-equating methods had less equating errors than the identity method. It can be seen in Figure 1 that as sample size grew beyond 50, the equating errors for the linear and LLC3 equipercentile methods were less than that for the identity method. The circle-arc method had the lowest equating

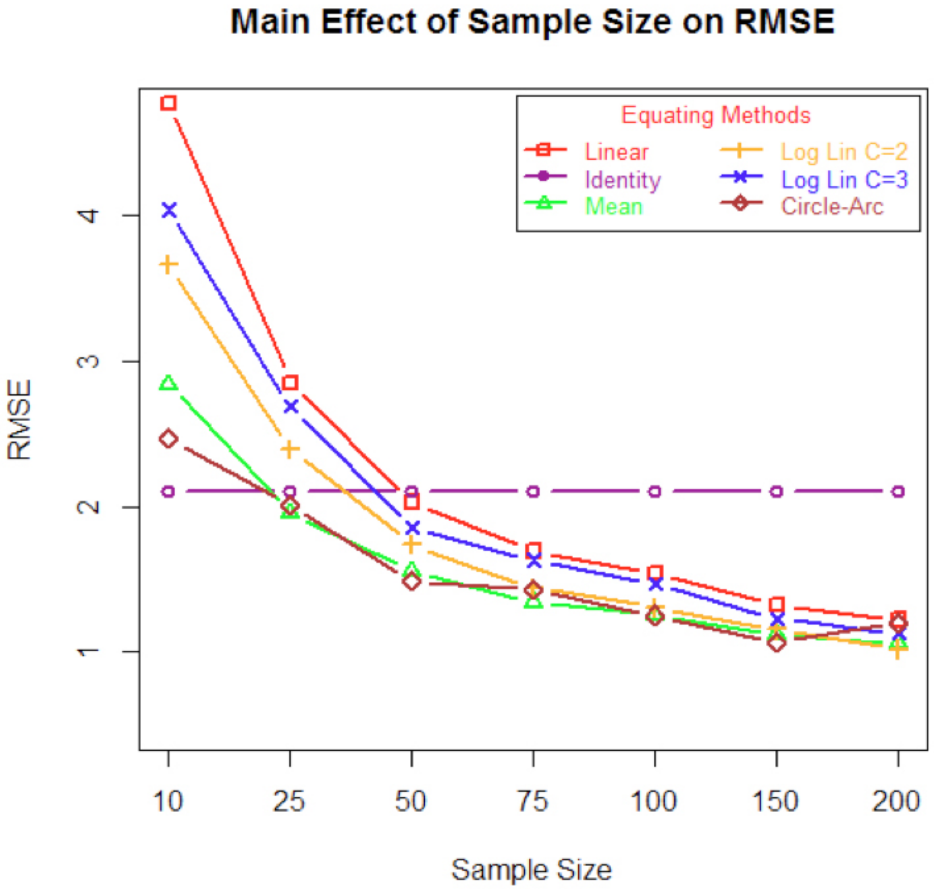

Figure 1. RMSE values for equating methods across different sample sizes.

error between sample sizes of 25 and 100. The mean-equating method produced the lowest RMSE for sample sizes between 100 and 200. At a sample size of 200, the LLC2 equipercentile-equating method had the lowest equating error compared to the other methods.

The Main Effect of Average Difference in Form Difficulty on Equating Error

The RMSE values for equating methods across various mean differences in form difficulty are shown graphically in Figure 2.

That the RMSE for identity method rose steadily with an increase in difficulty difference can be seen in this figure. However, the RMSE for other methods showed minimal variations. When the test forms differed by 0.1 SMD, the identity method had the lowest RMSE value. The RMSE for the circle-arc method decreased as the difference in difficulty between forms increased from 0.1 SMD to 0.4 SMD. When the test forms' difficulty differed by 0.4 SMD, the circle-arc method had

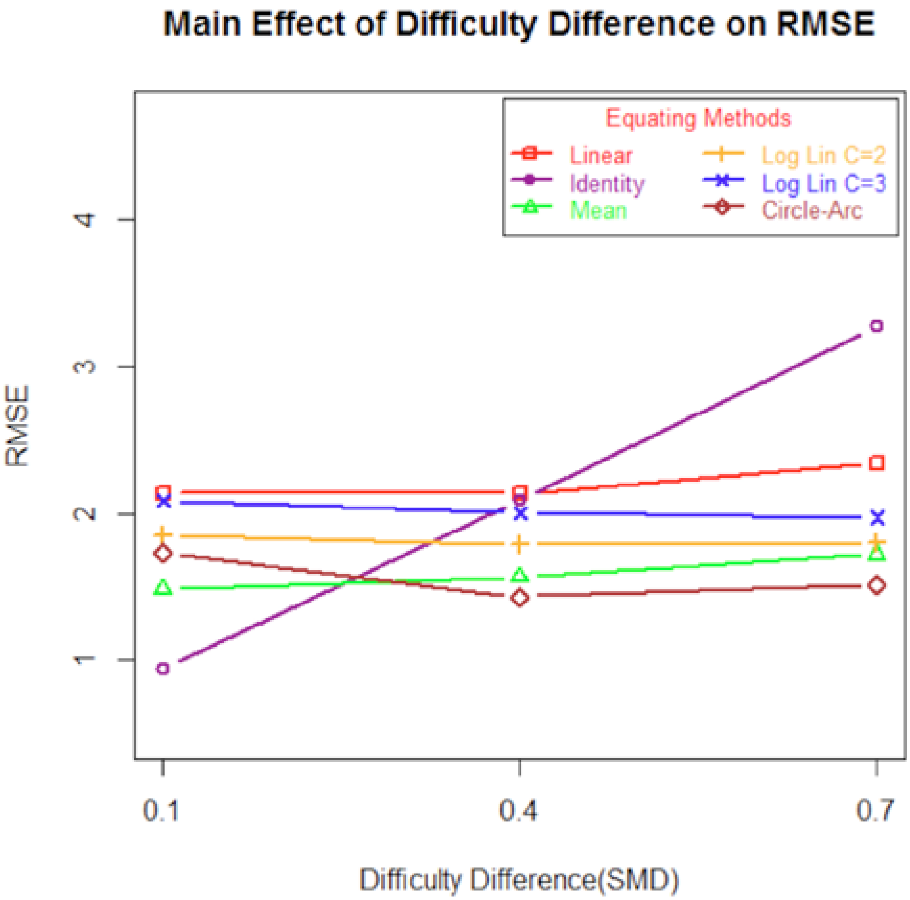

Figure 2. RMSE values of equating methods for various mean differences in form difficulty.

the lowest RMSE. The identity method had the greatest RMSE and the circle-arc method produced the smallest RMSE when the mean difference in form difficulty was between 0.4 and 0.7 SMD.

The Main Effect of Guessing Parameters on Equating Error

The RMSE values for equating methods across various guessing parameters are shown graphically in Figure 3.

In Figure 3, as the guessing parameter increased from 0.0 to 0.10, a slow decline in equating error with the identity and circle-arc methods could be seen to produce the lowest RMSE. However, when the guessing parameter was between 0.20 and 0.25, the RMSE for the circle-arc method increased whereas the mean-equating method produced the lowest RMSE. The methods of linear equating and 2- and 3-moments log-linear pre-smoothing produced a greater RMSE than the identity method for guessing parameters of 0.25.

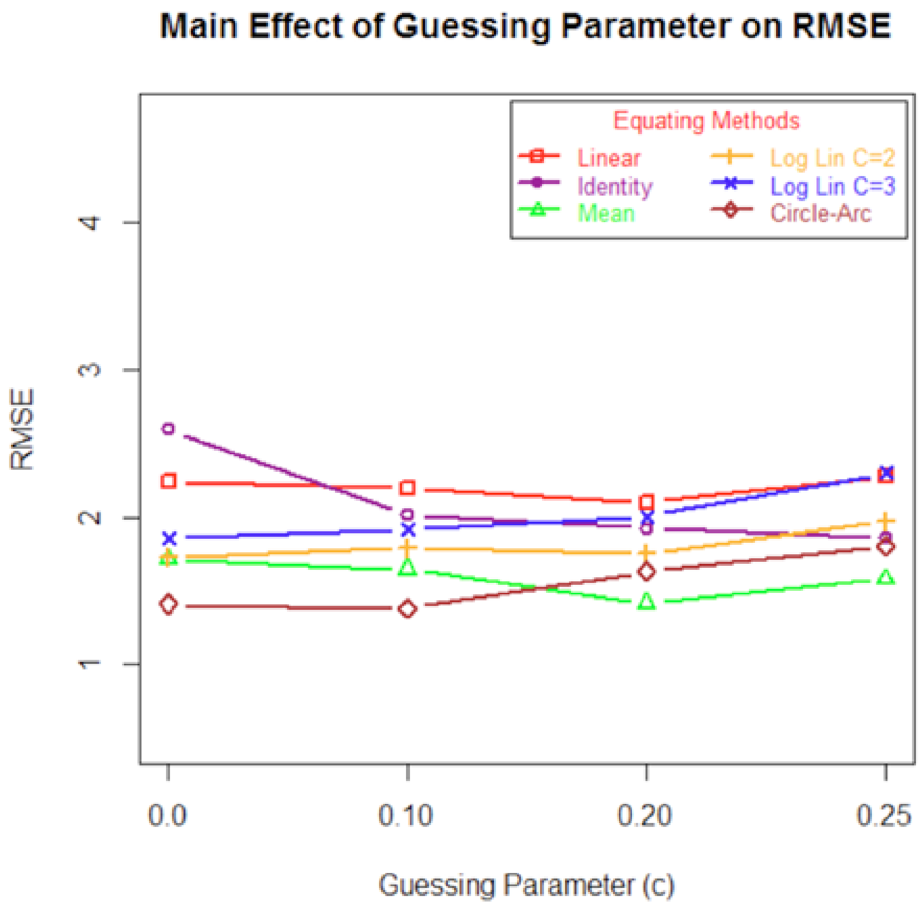

Figure 3. RMSE values for equating methods across different guessing parameters.

The RMSE varied across different guessing parameters for the equating methods. There was no significant variation on equation errors for any of the methods.

The Interaction Effects of Sample Size, Difference in Form Difficulty, and Guessing Parameters on the Equating Error (RMSE)

As is shown in Figure 4, the RMSE for all equating methods (except the identity method) decreased steadily. However, the RMSE for identity remained unchanged with an increase in sample size for all levels of difficulty difference and guessing parameters.

According to Figure 4, while there had been a steady increase in the RMSE for identity equating with an increase in difficulty difference between forms, no remarkable change in RMSE was seen with the other methods. The identity method produced the lowest RMSE for a difficulty difference of 0.1 SMD across all levels of sample size and guessing parameters. When difficulty difference between forms

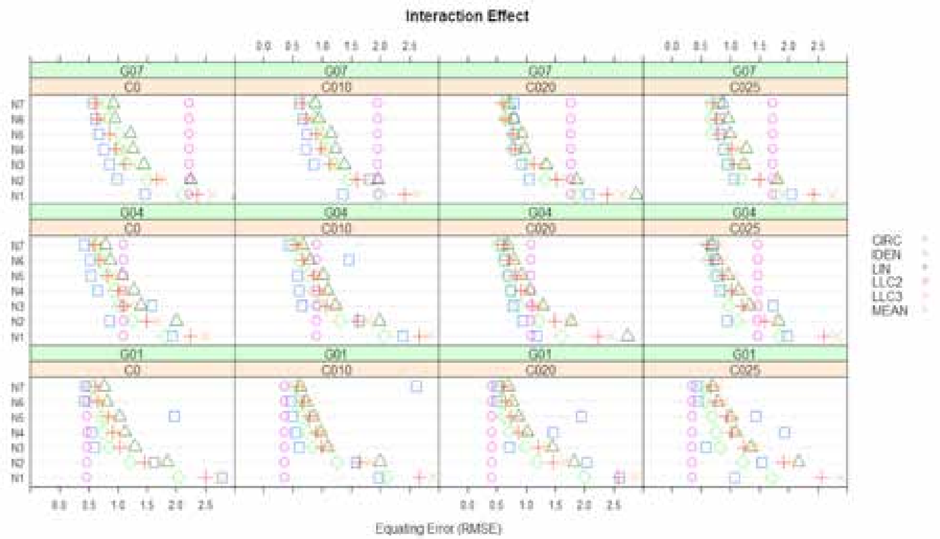

Figure 4. Interactive effects of sample size, difficulty difference between forms, and guessing parameters on equating error (RMSE).

was 0.4, the guessing parameter was 0.0 and 0.10, and when sample sizes were 100 or above, the identity method produced the greatest RMSE. In addition, the identity method had a higher RMSE than the other methods with sample sizes of 75 and above, difficulty differences of 0.4 SMD, and guessing parameters of 0.20 and 0.25. When the difference in form difficulty was too great (0.7 SMD), the identity method produced a larger RMSE than the other methods with sample sizes of 50 and above. As can be seen in Figure 4, the RMSE for the other methods did not change remarkably across different guessing parameters. However, when the guessing parameter was 0.10, the RMSE for the circle-arc method showed irregular variations.

Discussions

Evaluating the Main Effect of Sample Size on RMSE by Equating Method

The equating error for identity method clearly remained unchanged, whereas the equating errors for the other methods decreased steadily as sample size increased. The identity method produced the lowest RMSE with sample sizes between 10 and 25. When the sample size was 25, the equating errors for the methods of identity, circle-arc, and mean were nearly the same; the circle-arc and mean methods barely produced less error than the identity method. As can be seen in Figure 1, the equating error for the identity method was essentially independent of sample size, whereas the others were affected by sample size. In the literature, Skaggs (2005) found that forms should be equated with a sample size of 25. Furthermore, Livingston and Kim (2010) reported that the circle-arc method produced the lowest equating error when the sample size was 25. These results are identical with the findings from this study.

Hanson et al. (1994) claimed the minimum sample size required for equating to be 100 in his study, whereas Kim et al. (2008), and Kurtz and Dwyer (2013) suggested the required minimum sample size for equating to be 50. From Figure 1, all methods could be seen to produce a lower RMSE than the identity method for sample sizes of 50 or more. The results of the studies of Kim et al. (2008) and Kurtz and Dwyer (2013) confirm this finding. Equating forms have been said to reduce the equating error for sample sizes of 50 and above, so it has been suggested that forms should be equated regardless of equating method when the sample size is 50 or more. The circle-arc and mean methods produced more accurate results across sample sizes between 25 and 200. However, the LLC2 equipercentile equating method produced the most accurate results for a sample size of 200. Hence, the circle-arc and mean methods are preferable for sample sizes between 25 and 200, and the LLC2 equipercentile method is preferable for a sample size of 200. Kolen and Brennan (2004) suggested that the equipercentile equating method produced more accurate results with sample sizes of 1,500 or greater. Livingston (1993), Parshall et al. (1995), Skaggs (2005), and Livingston and Kim (2009b; 2010) conducted the log-linear pre-smoothing method at different moments in their studies and generally agreed that pre-smoothing substantially diminished the equating error (RMSE). These results are considered akin to the results of this study.

Evaluating the Main Effect of Average Difference in Form Difficulty on the RMSE by Equating Method

Based on Figure 2, differentiating the differences in form difficulty caused an increase in RMSE for the identity method. On the other hand, RMSE values for other methods remained unchanged. This finding is similar to the findings from Babcock et al. (2012). Harris and Crouse (1993) and Kolen and Brennan (2004) pointed out that not equating the forms (identity method) produced a smaller equating error as the difference between form difficulty lessened. Similarly, Babcock et al. (2012) also remarked that when the differences in form difficulty were similar, not equating the forms produced more accurate results. In this study, the identity method was concluded to be preferable only when test forms were nearly equivalent in difficulty. This finding is consistent with the results of the studies of Harris and Crouse (1993), Kolen and Brennan (2004), and Babcock et al. (2012). Since the RMSE for the identity method increases rapidly with an increase in difference in test form difficulty, test forms should be equated. Heh (2007) reported that when difficulty differences between forms are 0.3 SMD or greater, it becomes necessary to equate them. These results are identical to the findings from this study. Among the equating methods with respect to the RMSE, the circle-arc and mean-equating methods can be said to be the most preferable methods when the difference in form difficulty is too great. Babcock et al. (2012). suggested preferring the mean-equating method when test forms have substantially different difficulty levels and the examinees are equivalent in ability. These results are consistent with the findings of this study.

Evaluating the Main Effect of Guessing Parameters on the RMSE by Equating Method

Bozdağ (2010) concluded that equating test scores in terms of the probability of success and the probability of failing had an effect on the equating errors of equating methods. Overall, when increasing the level of guessing parameters, the RMSE values for each equating method varied. As shown in Figure 3, the circle-arc method was preferable for guessing parameters of 0.0 and 0.10, while the mean-equating method was preferable for guessing parameters of 0.20 and 0.25. The circle-arc method was the most affected method with respect to guessing parameters. The circle-arc method had an equating curve with a starting point, an endpoint, and an empirically calculated middle point. The lower endpoint of this curve was the lowest meaningful test score from the guessing parameter, and the upper endpoint was the maximum possible score. Guessing parameters directly affect the circle-arc method because guessing parameters determine the lower endpoint mathematically. Thus the reason for the effect of guessing parameters on the circle-arc method can be inferred as a result of their mathematical function. In brief, guessing parameters do not have a consistent effect on equating errors for forms.

Evaluating the Interactive Effect of Various Factors on the RMSE by Equating Method

As sample size increased, the RMSE for the identity method remained unchanged, while the other methods showed a steady decline in the RMSE with an increase in sample size at all levels for the other factors. Overall, this is completely consistent with the principle that increasing sample size decreases the equating error (Kolen & Brennan, 2004).

An increase in difference between form difficulty levels led to an increase in the RMSE for the identity method. The RMSE for equating methods, aside from the identity method, remained unchanged with differing difficulty levels between forms. It is clear from Figure 4 that the RMSE for the identity method was the most affected in terms of difference of difficulty levels between forms. Equating or not equating is decided by comparing the RMSE for the identity method with the other methods. If the RMSE for other methods is smaller than the RMSE for the identity method, equating should be conducted. Otherwise, there is no need to equate the forms. From Figure 4, the identity method may be concluded to not produce a significant equating error when the difficulty difference between forms is small; hence, the identity method provides the most accurate results in this situation. This finding is consistent with what has been reported in the literature (see Babcock et al., 2012; Devdass, 2011; Heh, 2007). The circle-arc method should not be preferred for equating because of fluctuations in its RMSE values when test forms have very similar difficulty levels. A difference of 0.4 SMD between forms produced significantly larger equating errors compared to a difference of 0.1 SMD. Thus the error for the identity method can be concluded to increase while the other methods remain unchanged. When difficulty difference is 0.7 SMD, all methods perform better than the identity method for sample sizes of 100 and above. The circle-arc, mean, and LLC2 equipercentile methods are preferred for sample sizes between 50 and 100 at all levels of the guessing parameter. When the difference in test difficulty was 0.7 SMD, identity equating had the largest equating error for sample sizes of 50 or more. Therefore, equating should be conducted when the sample size is 50 or more and the difficulty difference is 0.7 SMD. The circle-arc method is preferred for sample sizes of 50 and above and the guessing parameter is between 0.0 and 0.10. When the guessing parameter is between 0.20 and 0.25, the mean and LLC2 equipercentile methods provide less equating errors than the other ones. In general, any method other than identity equating becomes preferred when there are substantial differences between the forms' difficulty levels. Furthermore, as mentioned earlier, when test forms have similar difficulty levels, equating should not be conducted.

Differences in guessing parameters irregularly affected the various methods' equating errors. The circle-arc method was the most affected method for different guessing parameters. By increasing the guessing parameter, the equating errors for the circle-arc method fluctuated, especially when the difficulty differences between forms was similar.

When SMD is 0.1 and the parameter estimate is 0.0, the circle-arc method provided the most accurate results for sample sizes of 150 and 200. For this reason, the circle-arc method is more preferable under these situations. Compared to other methods, the identity method can be conducted to equate across all levels of sample sizes and guessing parameters except for 0.0 when the difficulty difference is 0.1 SMD. This confirms the suggestions of Skaggs (2005) and Heh (2007). As expected, identity equating is the most accurate when forms have similar difficulty levels. These results are parallel with the results from the studies of Harris and Crouse (1993) and Kolen and Brennan (2004).

When SMD is 0.4 and the guessing parameter is 0.0 or 0.20, the required minimum sample size to equate the forms is 25 for the circle-arc method, 50 for mean equating, 75 for LLC2 equipercentile equating, and 100 for the LLC3 equipercentile equating and linear-equating methods. All equating methods produce more accurate results at sample sizes of 100 or above across all levels of the other factors. When the guessing parameter is 0.10 and difficulty difference is 0.4 SMD, the circle-arc method reveals irregular variations with respect to RMSE. Under these circumstances, as only the circle-arc method is preferable at a sample size of 50, all methods can be preferred at sample sizes of 150 and above. When SMD is 0.4 and the guessing parameter is 0.25, all equating methods are better than the identity method with respect to RMSE at sample sizes of 75 or more, so equating should be conducted under these situations. This finding is consistent with the findings of other researches (see Heh, 2007; Skaggs, 2005).

When the SMD is 0.7, the circle-arc and mean-equating methods are the most accurate methods across all sample sizes and all levels of guessing parameters except for 0.25. The linear-equating method can be used for sample sizes of 50 and above, whereas other methods are used for sample sizes of 25 and above. In these situations, this is in place of identity equating. When SMD is 0.7 and the guessing parameter is 0.25, the results are similar to before. However, the methods of linear-equating and LLC3 equipercentile equating can be preferred for sample sizes of 50 and above, and the others are preferred for samples sizes of 25 and above. Consequently, the difference in form difficulty requires equating test forms when the difference in test form difficulty is large. These results are identical with the findings from Babcock et al. (2012).

Conclusion and Recommendations

Sample size, as well as other factors, has a significant effect on equating test scores. If a sample is large and representative, equating in a sample may accurately represent the population. Large samples (i.e. 1,500 and above) are recommended for test equating. As mentioned earlier, for different reasons this may not always be possible. In the test equating process when sample size is small enough, the sample representation of the population is less and this leads to the occurrence of equating errors. The main aims of this study have been to decide whether equating is required or not, to decide which methods produce more accurate results, and to decide which are preferred with respect to RMSE if equating should be required.

In the study, the main and interactive effects of the equating methods were examined separately. One conclusion from this study which is consistent with other research is that increasing sample sizes reduces the standard equating error and hence the RMSE for equating methods. However it does not have an effect on the identity method. Overall, an increase in sample size leads to a decrease in equating error for all except the identity method. It can be concluded that when a sample size is 50 or more, equating test forms produces more accurate results than not equating. Therefore, it is recommended that forms should be equated for sample sizes of 50 and above regardless of equating method. The circle-arc and mean-equating methods can be preferred even for sample sizes of 25. When the sample size is extremely small (such as 10), equating is not an essential requirement. However, equating may be beneficial across these sample sizes only if the difficulty difference is 0.4 SMD or greater.

When test forms have similar difficulty levels (0.1 SMD), it is not required to equate the test forms. However, when the sample sizes are increased to 150 or more, the circle-arc equating method provides the least amount of errors, so it may be preferable. As the difficulty difference increases up to 0.4 SMD, the identity method produces more equating errors, thus necessitating equating. When the equated forms have an extreme difference in difficulty (0.7 SMD), equating should be conducted.

Guessing parameters also affect the methods' equating errors. However, unlike the others, this effect is not steady. Equating test scores through probability of failure provides more accurate results. The circle-arc method produces less equating errors than the others when equating test scores through probability of failure. Hence, it is more preferable in this situation. However, when guessing parameters are manipulated, the equating error for the identity method decreases. In this situation, it is beneficial to check Table 2 for deciding whether equating is required or not and which method may be preferred.

According to the results of this study, Table 2 (shown below) is suggested to practitioners as a guide for studying with small-sample equating methods using the random-group design. The rows on this table represent various sample sizes. The columns represents the average difficulty difference between forms (upper row) and guessing parameters (lower row). Lastly, each cell in the table represents the equating method which is recommended at those levels per the factors.

Table 2

Recommended Equating Methods for Random Group Design Using Small Samples at All Levels and for All Factors in the Study

b		0.1				0.4				0.7			
	C	0.0	0.10	0.20	0.25	0.0	0.10	0.20	0.25	0.0	0.10	0.20	0.25
N	200	C	ID	ID	ID	All	All	All	All	All	All	All	All
	150	C	ID	ID	ID	All	All	All	All	All	All	All	All
	100	ID	ID	ID	ID	All	C-ME-LLC2	All	All	All	All	All	All
	75	ID	ID	ID	ID	C-ME-LLC2	C-ME	All	All	All	All	All	All
	50	ID	ID	ID	ID	ME	C	C-ME	All	All	All	All	All
	25	ID	ID	ID	ID	C	ID	C	C-ME	C-ME-LLC2-LLC3	C-ME-LLC2-LLC3	C-ME-LLC2-LLC3	C-ME-LLC2
	10	ID	ID	ID	ID	ID	ID	ID	ID	C-M	C	ID	ID

Note. C: Circle-Arc, ID: Identity Equating, ME: Mean Equating, LIN: Linear Equating, LLC2: 2-moments log-linear pre-smoothing method, LLC3: 3-moments log-linear pre-smoothing method; ALL: All methods except for identity equating.

In this study, some factors such as sample size, difficulty difference, and guessing parameters were used at various levels to examine the effect of these factors on the accuracy of equating methods. Sample sizes varied from 10 to 200, test forms differed in difficulty at 0.1, 0.4, and 0.7 SMD, and guessing parameters were manipulated at 0.0, 0.10, 0.20, and 0.25. The same study could be repeated at different levels for these factors. Moreover, different factors such as item discrimination, test length, ability distribution, and cut-off scores could be included in further studies. New equating methods such as synthetic function (Kim et al., 2008) and nominal-weights mean equating (Babcock et al., 2012), or IRT-based equating methods (item response theory) could be conducted and compared with other methods. Simulated data was produced for this study. This study could be repeated with real data. Furthermore, similar studies could be made using the NEAT design. In this study, equating methods were compared only in terms of the RMSE statistic. In further studies, standard equating error and equating bias could be used to examine the accuracy of equating methods.

References

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

Babcock, B., Albano, A., & Raymond, M. (2012). Nominal weights mean equating: A method for very small samples. *Educational and Psychological Measurement, 72*(4), 608–628.

Bozdağ, S. (2010). Şans *başarısının test eşitlemeye etkisi* [The effect of chance of success on test equalization] (Master's thesis, Mersin University, Mersin, Turkey). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

Cui, Z., & Kolen, M. J. (2009). Evaluation of two new smoothing methods in equating: The cubic b-spline pre-smoothing method and the direct pre-smoothing method. *Journal of Educational Measurement, 46*(2), 135–158.

Devdass, S. (2011). *Conditions affecting the accuracy of classical equating methods for small samples under the NEAT design: A simulation study* (Doctoral dissertation). University of North Carolina, NC.

Donlon, T. (Ed.). (1984). *The College Board technical handbook for the scholastic aptitude test and achievement tests*. New York: College Entrance Examination Board.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). *Principles and practices of test score equating* (ETS Research Report No. RR-10-29). Princeton, NJ: ETS.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and Applications.* New York, NY: Springer.

Hanson, B. A., Zeng, L., & Colton, D. (1994). *A comparison of pre-smoothing and post-smoothing methods in equipercentile equating* (ACT Research Report No. 94-4). Iowa City, IA: American College Testing.

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education*, *6*, 195–240.

Heh, V. K. (2007). *Equating accuracy using small samples in the random groups design* (Doctoral dissertation). Patton College of Education at Ohio University, Athens, OH.

Kim, S., vonDavier, A. A., & Haberman, S. (2006). An alternative to equating with small samples in the non-equivalent groups anchor test design. Paper presented at the annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Kim, S., vonDavier, A. A., & Haberman, S. (2008). Small-sample equating using a synthetic linking function. *Journal of Educational Measurement, 45*(4), 325–342.

Kolen, M. J. (1988). Traditional equating methodology. *Educational Measurement: Issues and Practice, 7*(4), 29–37.

Kolen, M. J., & Brennan, R. J. (1995). *Test equating: Methods and practices*. New York, NY: Springer-Verlag.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Method and practice* (2nd ed.). New York, NY: Springer-Verlag.

Kurtz, A. M., & Dwyer, A. C. (2013). *Small sample equating: Best practices using a Sas Macro*. Retrieved from http://analytics.ncsu.edu/sesug/2013/BtB-11.pdf

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

Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: ETS.

Livingston, S. A., & Kim, S. (2009b). The circle-arc method for equating in small samples. *Journal of Educational Measurement, 46*(3), 330–343.

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Öztürk, N. (2010). *Akademik personel ve lisansüstü eğitimi giriş sınavı puanlarının eşitlenmesi üzerine bir çalışma* [A study on equating the scores of the academic staff and postgraduate education entrance exam] (Master's thesis, Hacettepe University, Ankara, Turkey). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/

Öztürk, N., & Anıl, D. (2012). Akademik personel ve lisansüstü eğitimi giriş sınavı puanlarının eşitlenmesi üzerine bir çalışma [A study on equating the scores of the academic staff and postgraduate education entrance exam]. *Eğitim ve Bilim, 37*(165), 181–193.

Parshall, C. G., Houghton, P. D. B., & Kromrey, J. D. (1995). Equating error and statistical bias in small sample linear equating. *Journal of Educational Measurement, 32*(1), 37–54.

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement, 42*(4), 309–330.

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