TURJE
www.turje.org

Volume 5 Issue 3

DOI: 10.19128/turje.16916

Turkish Journal of Education
Research Article on Educational Measurement and Assessment

Received 07.07.2016
Accepted 29.07.2016

Comparison of several small sample equating methods under the NEAT design

Serdar Caglak
Eskisehir Osmangazi University, Eskisehir, Turkey, scaglak@ogu.edu.tr

ABSTRACT The aim of this study is to compare the performances of Identity, Nominal Weights Mean (NMW), and Circle-Arc (CA) equating methods under the Non-Equivalent Groups Anchor-Test (NEAT) design. Synthetic equating functions (SFs) of the NWM and CA (NWS and CAS) were also created using an equal weighting system (w = 0.5). Different sizes of small examinee samples (n = 10, 20, 50, 100) were used to equate new test forms to base test forms. Chained Equipercentile (CE) with bivariate log-linear presmoothing was used as population criterion equating function to compare the performances of the equating methods. Overall, the identity (ID) equating was the most favorable, but the NWS method produced less equating error than the ID and Tucker Linear (TL) equating methods under specific simulation conditions. The use of the SF of the NWM method can be used in practice to equate the test forms with samples less than 25 examinees. In future studies, the SFs of other existing equating methods should be tested to determine the best performing equating method(s) for small sample equating.

Keywords *small samples, equating, synthetic functions, NEAT design.*

Küçük örneklemlerde kullanılan bazı test eşitleme yöntemlerinin DOOT deseni altında karşılaştırılması

ÖZ Bu çalışmanın amacı, Identity (İD), Nominal Weights Mean (NWM) ve Circle-Arc (CA) test eşitleme yöntemlerinin performanslarını denk olmayan gruplar ve ortak soru içeren test (DOOT) deseni altında karşılaştırmaktır. Bu yöntemlere ait yapay test eşitleme fonksiyonları (NWS ve CAS) eşit ağırlıklandırma sistemi (w = 0.5) kullanılarak ayrıca oluşturulmuş ve sonrasında orijinal test eşitleme yöntemleriyle (İD, NW ve CA) karşılaştırılmıştır. Yeni test formlarını, referans test formlarına eşitlemek için farklı büyüklükte küçük örneklemler (n = 10, 20, 50, 100) kullanılmıştır. Chained Equipercentile (CE) yöntemi loglinear data düzeltme tekniği ile birlikte kullanılarak, evrendeki test formları arasında fonksiyonel bir ilişki kurulmuştur. Testler arasında kurulan bu fonksiyonel ilişki, İD, NWM, CA, NWS ve CAS test eşitleme yöntemlerinin performanslarını karşılaştırmada bir ölçüt olarak kullanılmıştır. Sonuç olarak, İD eşitleme yöntemi en uygun yöntem olarak tespit edilmiştir. Ancak NWS yöntemi de bazı durumlarda İD ve TL yöntemlerinden daha az hata üretmiştir. NWS yöntemi, örneklem sayısının 25`ten az olduğu durumlarda İD'ye alternatif olarak kullanılabilecek niteliktedir. Sonraki çalışmalarda ise, var olan diğer test eşitleme yöntemlerinin yapay fonksiyonları test edilmeli ve küçük örneklemler ile kullanılabilecek en uygun yöntem(ler) tespit edilmelidir.

Anahtar Kelimeler *küçük örneklemler, test eşitleme, yapay fonksiyonlar, DOOT deseni.*

GENİŞLETİLMİŞ ÖZET

Çoktan seçmeli testler bireyler hakkında önemli karar verme süreci içerisinde pratikte yaygın olarak kullanılmaktadır. Örneğin, adayların seçimi ve yerleştirilmesi, yükseköğretime kabulü, bir kuruma atanmaları ve bir işe uygunluklarının tespiti bunlardan bazılarıdır. Özellikle Amerika Birleşik Devleteri`nde yapılan öğretmen alan sınavlarında ve sertifika programlarında çoktan seçmeli testler sıklıkla kullanılmaktadır. Bu testleri kullanmadaki amaç her ne kadar güvenilir ve geçerli sonuçlar elde etmekse de, uygulamadaki en büyük problemlerden birisi testlere ait materyallerin güvenliğinin ve gizliliğinin sağlanamamasıdır. Bu problemi aşmak için testler farklı zamanlarda uygulanmakta ancak bu durum da farklı problemlerin ortaya çıkmasına zemin oluşturmaktadır. Örneğin, faklı zamanlarda adaylara uygulanan sınavların zorluk dereceleri farklılık göstermekte, bu da adayların test puanlarının adaletli bir şekilde karşılaştırılmasının önüne geçmektedir. Bu sebeple, testlerin zorluk derecelerinin eşitlenmesi ve sonrasında eşitlenmiş test puanlarının adaylara duyurulması gerekmektir. Testler arasındaki ortalama zorluk derecelerinin istatiksel olarak eşitlenmesi işlemi "test eşitleme" olarak adlandırılmaktadır (Kolen ve Brennan, 2004, s. 2). Test eşitleme işleminin tam olarak gerçekleştirilmesi, sınavları uygulayan kurumların her zaman arzu ettiği bir durumdur. Ancak, testi alan aday sayısının az olması ve/veya testin yapısına uygun olmayan test eşitleme yöntemlerinin kullanması, yanlış test puanlarının ortaya çıkmasına ve dolayısıyla da adayların yanlış değerlendirilmesi gibi problemlere yol açmaktadır (Caglak, 2015).

Sınavda başarılı olan adayların doğru tespit edilmesi gerekmektedir. Ancak psikometri uzmanları, testleri eşitlerken kullanacakları örneklem sayısının yeterli olmaması durumu ile karşılaşabilmektedir. Ayrıca, uzmanlar bu gibi durumlarda çoğunlukla nasıl hareket edeceklerini de bilememektedirler. Genel olarak, pratikte iki durum ile karşılaşılmaktadır; 1) sistematik hata riskini göze alıp, testleri eşitlemeden adayların test puanları açıklamak ya da 2) testlerin küçük örneklemlerle eşitlenmesiyle oluşan rastgele hata miktarını göz ardı etmek (Caglak, 2015). Diğer bir deyişle, alan uzmanlarının aslında fazla seçeneği yoktur. Çünkü her iki durumda da hata riski çok yüksektir. Anlamlı ve karşılaştırılabilir test puanlarını elde etmek için uzmanlar uygun yöntemi seçmelidirler. Kısaca, kullanılacak yöntemin ne çok fazla sistematik hata ne de çok miktarda rastgele hata üretmesi beklenir.

Bazı araştırmacılar küçük örneklemler ile kullanılacak eşitleme yöntemlerini test etmişler ve bulgularına göre ya yeni yöntemler önermişler ya da var olan yöntemleri birleştirerek örneklem sayısının azlığından meydana gelebilecek problemlerin önüne geçmeye çalışmışlardır. Örneğin, Harris (1993) identity (İD) yöntemini örneklem sayısının çok az olduğu durumlar için önermiştir. Aslında bu yöntem farklı test kitapçığından elde edilen puanları kendi orijinal ölçeklerinde değerlendirilmekte ve bu puanları direkt olarak eşitlenmiş gibi kabul etmektedir. Kolen ve Brennan (2004) İD yöntemin psikometrik açıdan bire bir aynı olan testlerin eşitlenmesi için uygun olduğunu ifade etmişlerdir. Fakat bu duruma pratikte fazla rastlanılmamaktadır. Çünkü farklı zamanlarda uygulanan sınavlar farklı sorular içerdiğinden, testlerin zorluk derecelerinin aynı olması beklenemez. Babcock ve arkadaşları (2012), Nominal Weights Mean (NWM) adında pratikte çok kullanılan Tucker Linear (TL) yönteminin daha basit bir versiyonunu önermişlerdir. Bu yeni yöntem, TL'nin genel matematiksel eşitliğindeki bazı parametrelerin küçük örneklemlerle doğru kestirilemeyeceğini göstermektedir. Livingston ve Kim (2008) var olan test eşitleme yöntemlerinin çok düşük ve çok yüksek olan test puanları eşitleme de yetersiz kaldığını öne sürmüş ve bu duruma çözüm oluşturmak için Circle-Arc (CA) adında yeni bir yöntem önermişlerdir. Kim ve arkadaşları (2008), İD metodunun küçük örneklemlerle ürettiği sistematik hatayı azaltmak için İD'nin diğer test eşitleme yöntemleri ile birleştirilmesi gerektiğini ifade etmişlerdir. Bu sebeple Kim ve arkadaşları (2008), genel matematiksel bir eşitlik ortaya atmışlar ve bu eşitliğe yapay test eşitleme fonksiyonu adını vermişlerdir. Bu fonksiyonun avantajı, herhangi bir test eşitleme yönteminin bir ağırlıklandırma sistemi aracılığıyla İD ile birleştirilebilir olmasıdır.

Literatürde test eşitleme ile ilgili birçok çalışma mevcuttur. Ancak küçük örneklemler ile yapılan çalışma sayısı çok azdır. Ayrıca, bu yöntemlere (CA ve NWM) ait yapay test eşitleme fonksiyonlarını (CAS ve NWS) inceleyen başka bir çalışmaya literatürde rastlanılmamıştır. Bu çalışmanın diğer çalışmalardan farkı İD yöntemine alternatif bir çözüm önermesidir. Bu çalışmada önerilen yapay test eşitleme fonksiyonları, özellikle öğrenci sayısının az olduğu durumlarda rahatlıkla kullanılabilecek niteliktedir. Pratikte sıklıkla kullanılan vize ve final sınavları, ders içi değerlendirme faaliyetlerinde kullanılan quizler bu gibi durumlara örnek olarak verilebilir.

Önerilen yöntemlerin özellikleri dikkate alınacak olursa, bu yöntemler örneklem sayısının azlığından kaynaklanan problemlere çözüm olabilecek potansiyele sahiptirler. Bu sebeple, NWS ve CAS'ın İD'den daha az hata üretmesi beklenmektedir. Genel olarak çoktan seçmeli ve şans başarısının var olduğu DOOT deseni altında uygulanan sınavlarda gözlenen, doğru ve yanlış cevapları elde etmek için, bu çalışmada üç parametreli Madde Tepki Kuramı (MTK) ile bilgisayar ortamında öğrenci verileri oluşturulmuştur. Belirtilen test eşitleme yöntemleri ve örneklem sayılarına ek olarak, testlerin zorluk dereceleri ile öğrenci gruplarının kabiliyet düzeyleri göz önünde bulundurulmuştur. Buna bağlı olarak da 64 farklı simülasyon durumu oluşturulmuştur. Örneklem sayısı yeni test için 10, 20, 50 ve 100 olarak belirlenmiştir. Ancak, bu sayı referans testi (ilk test) için 200 olarak sabit tutulmuştur. Test eşitleme işlemi her bir durum için 1000 defa tekrarlanmış ve test eşitleme yöntemleri ürettikleri hatalar bakımından karşılaştırılmıştır.

Bulgulara göre, testlerin zorluk derecelerinin ve testi alan grupların kabiliyet düzeylerinin birbirine yakın olduğu durumlarda, NWS İD'den daha az hata üretmiştir. Ancak aksi durumlarda İD yöntemini kullanmak, ürettiği hata açısından NWS yöntemine göre daha avantajlıdır. Test edilen yöntemler simülasyon durumunun özelliklerine göre, ürettikleri hata bakımdan değişkenlik göstermektedir. Ürettikleri toplam hata bakımından, yapay fonksiyonlar orijinal yöntemlerden genel olarak daha az hata üretmiştir. Sonuç olarak, küçük örneklemlerde kullanılacak uygun yöntemin belirlenmesinde iki temel unsur dikkat çekmektedir. Bunlar, toplam test puanları arasındaki standartlaştırılmış ortalama fark ile ortak testlerden alınan test puanları arasındaki standartlaştırılmış ortalama farktır. Eğer bu iki fark çok büyük ise, testleri eşitlemek daha fazla hataya yol açmaktadır. Bu iki farkın küçük olduğu durumlarda ise, NWS'nin İD'ye göre daha az hata üretmektedir.

INTRODUCTION

Standardized tests are widely used in practice for different purposes as a part of high-stakes decision making process, such as in selection, admission, qualification, certification, placement, employment, and so on. Even though the purpose of standardized testing is to provide reliable and valid assessments, one of the practical problems is to maintain the test security and confidentiality of testing materials in almost any examination program. Elimination of any potential threat to those aspects of standardized testing is very crucial in order to obtain meaningful and fair results. Multiple versions of tests are therefore used to prevent any threat to security and confidentiality of tests. However, their usage causes some other problems in testing. Inaccurate score comparability, for example, becomes an outstanding practical problem since test forms show different psychometric characteristics when multiple test forms are used. Especially, average difficulty levels of tests may show substantial differences across multiple versions since identical psychometric characteristics are not always observed, which simply pose a validity threat to examinees' test scores. For a meaningful and accurate score comparability among test takers, test form difficulty differences must be adjusted before releasing examinees' test scores. The statistical procedure that is used to adjust test form difficulty differences is called "equating" (Kolen & Brennan, 2004, p.2). Successful and accurate equating is always desired in any examination program; however, existence of sampling fluctuations and inappropriate use of equating methods may yield inequivalent or incompatible test scores.

True values of parameters at the population level are typically unknown. Sample statistics are therefore used to make inferences about those parameters. It is very important for practitioners to know whether randomly drawn samples are representative of their populations in order to make correct and defensible statements about parameters of interest (Peterson, 2007). However, randomly drawn samples always vary, and thus, sample statistics (e.g., mean, standard deviation, etc.) obtained from those samples differ due to sampling fluctuations or sampling errors (Howell, 2007). Like many other statistical procedures, test score equating is also subject to sampling variability (Livingston, 1993). Substantial change in size of samples therefore affects the functional relationship or equating relationship between the new test form and the base test form of tests (Kolen & Brennan, 2004; Livingston & Kim, 2011). In other words, the estimated equating relationship can accurately represent the equating relationship at the population level if the randomly drawn sample data is large enough. Otherwise, the equation function between the test forms may considerably differ from that of the population (Kim, von Davier, & Haberman, 2006).

Small sample equating typically occurs in teacher certification/licensure examinations due to the low number of teacher candidates taking specific subject area tests (Kim et al., 2006). Compared to other examination programs, such as K-12 assessments, teacher certification examinations are administered to very small numbers of test takers (Babcock, Albano & Raymond, 2012; Kim, Livingston, & Lewis, 2011). Since timely score release has priority among the other operational tasks in teacher licensure programs, there is no possibility of collecting more data to equate test forms using larger samples (Kim & Livingston, 2010). Moreover, test items are replaced with new test items periodically due to item exposure in many testing programs. Examinees also respond to both specific unique items and common set of items in both test forms under the non-equivalent groups anchor test design (NEAT). However, some degree of inaccuracy still exists even though the common set of items in both test forms is used to establish a functional relationship between the test forms (Livingston, 1993). Small sample test equating, therefore, becomes an avoidable situation in most teacher certification examinations.

Since fail or pass status is desired for each of the teacher candidates, correct classification of the candidates using their test scores must be the main goal of any teacher examination program. However, practitioners have difficulties to decide what to do when the number of examinees taking the teacher certification examinations is very small because the size of the samples is very crucial for obtaining accurate and comparable test scores. The absence of large data forces practitioners either not to equate test forms at the expense of getting a large equating bias or to equate the test forms at the expense of getting a large standard error (Caglak, 2015). In other words, practitioners do not have many choices to consider, and thus, the right choice of an appropriate equating method becomes a priority for the practitioners in order to obtain meaningful and comparable examinee scores across different test forms.

Error in Equating

Different test forms are connected to each other using a statistically obtained linking function or more specifically with an equating function. Some degree of error is therefore always present whenever test

forms are equated regardless of the size of samples used in this statistical procedure. Since no unique method exists in practice to equate test forms, the use of different equating methods may also yield obtaining different results or some degree of bias even if same examinee samples are used to equate the test forms. Kolen and Brennan (2004) describe two types of errors in the test score equating context: random error and systematic error.

Random error refers to sampling error or standard error of equating (SEE). Similarly, systematic error is attributed to equating bias (BIAS), which occurs due to violations of assumptions of test score equating or because of using inappropriate methods to equate the test forms (Kolen & Brennan, 2004). Briefly, sample fluctuations are the main factors affecting the magnitude of random error while the method related concerns or problems yield equating bias. Random error can be easily quantified or estimated using several statistical procedures (e.g., bootstrapping, delta4, etc.). However, it is very hard for practitioners to quantify the systematic error, but it can be controlled through a careful test development procedure and also with the use of appropriate data collection design and equating method (Kolen & Brennan, 2004).

Small Sample Equating Methods

Identity equating

When equating is unnecessary or unwarranted, the use of the ID equating is recommended (Harris, 1993). The ID equating refers to no equating since the slope of the equating function is specified as 1 and the intercept is fixed to 0. In other words, scores in the scale of the new test form are transformed to the scores in the scale of the base or reference test form. However, the transformed scores are still equivalent to the scores in the original scale of the new test form. Equation 1 shows this functional relationship between the test forms, where x is a randomly observed score on the scale of Form X (new test form) and e_y(x) is its equivalent on the scale of Form Y (base test form).

$$e_y(x) = 1*x + 0 \tag{1}$$

Since the scores are still assumed to be equivalent to the scores in their original scale, the standard error equating is zero by definition in the ID equating (Kim et al., 2011). The use of ID equating in practice is therefore often recommended when the number of new form test takers is less than 100 (Kolen & Brennan, 2004). However, its use causes substantial equating bias especially when the psychometric characteristics of the test forms are different (Kim et al., 2011).

Nominal weights mean equating

Babcock et al. (2012) introduced a simplified version of Tucker Linear (TL) equating, which is called Nominal Weights Mean (NWM) equating. Nominal weights are used to replace the variance and covariance terms in the TL equating with the numbers of total and anchor items and also with the numbers of examinees taking the test. The purpose of this replacement is to simplify the equating function due to the fact that the variance and covariance terms are not accurately estimated when the sample size is small (Babcock et al., 2012). In NWM equating, the standard deviations of scores on both Form X and Form Y are also assumed to be equal, and thus, the synthetic means in the TL equating are transformed to their simplified versions as shown in equation 2 through 9. More details about the TL equating can be found in Kolen and Brennan (2004). The advantage of using NWM equating arises when the number of new form test takers is relatively small (such as, 10, 20, and 50) and also when there is a high risk of obtaining large equating bias from traditional equating methods (e.g., Tucker Linear, Chained Equipercentile, etc.).

$$e_y(x) = \frac{\sigma_s(Y)}{\sigma_s(X)}[x - \mu_s(X)] + \mu_s(Y) \tag{2}$$

$$e_y(x) = x - \mu_s(X) + \mu_s(Y) \tag{3}$$

$$\mu_s(Y) = \mu(Y) + w_X\gamma_Y[\mu_{XV} - \mu_{YV}] \tag{4}$$

$$\mu_s(X) = \mu(X) - w_Y\gamma_X[\mu_{XV} - \mu_{YV}] \tag{5}$$

$$\gamma_Y = \frac{\sigma_{YV}}{\sigma_V^2} \text{ and } \gamma_X = \frac{\sigma_{XV}}{\sigma_V^2} \quad (6)$$

In NWM equating, the γ terms in equations 4 and 5 are replaced with the ratio of the total test length $(K_{X;Y})$ to the anchor test length (K_V) in the both test forms. Also, w represents the ratio of the form specific sample size over the total sample size of the examinees taking both the Form X and Form Y as shown below:

$$w_X = \frac{N_X}{N_X + N_Y} \text{ and } w_Y = \frac{N_Y}{N_X + N_Y}$$

When γ and w terms are replaced with N and K terms, the synthetic means in TL equating become

$$\mu_S(Y) = \mu(Y) + \frac{N_X}{N_X + N_Y}\frac{K_Y}{K_V}[\mu_{XV} - \mu_{YV}] \quad (7)$$

$$\mu_S(X) = \mu(X) + \frac{N_Y}{N_X + N_Y}\frac{K_X}{K_V}[\mu_{XV} - \mu_{YV}] \quad (8)$$

After all of these equations are put together, the equating function of NWM equating takes its final form as shown in equation 9.

$$e_y(x) = x - \mu(X) + \mu(Y) + \left[\frac{N_X K_Y + N_Y K_X}{[N_Y + N_X]K_V}\right][\mu_{XV} - \mu_{YV}] \quad (9)$$

Circle-arc methods of equating

Livingston and Kim (2008) introduced two versions of Circle-Arc (CA) methods of equating (symmetric and simplified) to establish a function between the test forms especially when the sample size of test takers is less than 30 for the new test form. The difference between the two versions comes from how the equating function is constructed. As seen on the left hand side in Figure 1, in the symmetric version, the equating function passes through three pre-specified points. However, the equating curve is divided into two parts in the simplified version: the linear component $L(x)$ that connects the two pre-specified end points and the curvilinear component that deviates from the line connecting the two end-points.

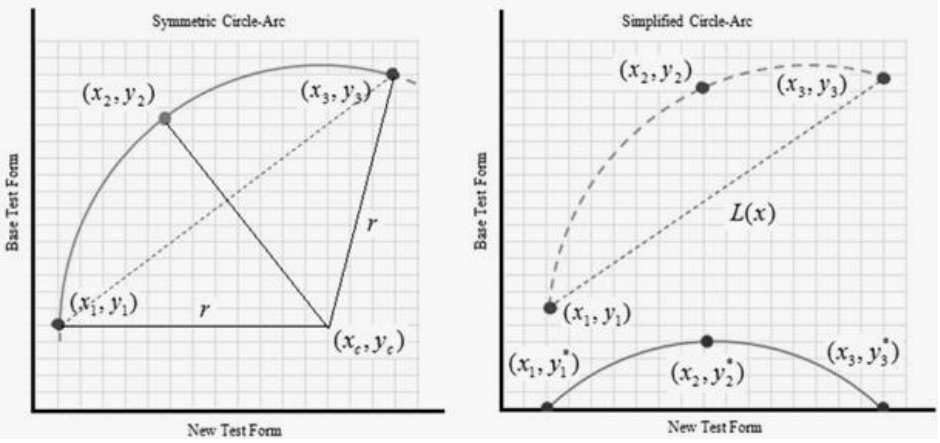


Figure 1. Symmetric and Simplified Versions of Circle-Arc Method of Equating

The pre-specified end points represent the possible minimum and maximum scores on the test, but the middle point is determined using mean test scores on both the new and base test forms. Depending on the data collection design, the equating method that is used to estimate the middle point varies. Traditional mean or linear equating, for example, is preferred if the data is collected through a random

groups test design; otherwise, the practitioners prefer using chained means or chained linear methods when a non-equivalent groups anchor test design is employed to collect examinee data (Babcock et al., 2012).

Simple mathematical calculations may require obtaining equating functions for both versions, but the estimation of the geometric projection of the curve on the x-axis in the simplified version is actually a more complex procedure while an arc of a circle is just fitted to three points in the symmetric version. Livingston and Kim (2011) pointed out that both versions produce identical results. Hence, in this study, the symmetric version was compared with the other chosen equating methods due to its mathematical and conceptual simplicity. To gain more in-depth understanding of the symmetric CA method, let's label (x_1, y_1), (x_2, y_2), and (x_3, y_3) as the pre-specified points as shown on the left hand side in Figure 1. The x-axis represents the scores on the Form X and the y-axis represents the scores on the Form Y. The lowest and highest possible scores on the both test forms are expressed by (x_1, y_1), and (x_3, y_3) points. The equating curve with a radius *r* and center of (x_c, y_c) passes through (x_2, y_2) to connect (x_1, y_1) with (x_3, y_3). Equation 10 is used as the equating function if (x_2, y_2) is below the straight line connecting (x_1, y_1) to (x_3, y_3). Otherwise, equation 11 is used as the equating function to transform Form X scores to the scores on the scale of Form Y.

$$e_y(x) = y_c - \sqrt{r^2 - (X - x_c)^2} \tag{10}$$

$$e_y(x) = y_c + \sqrt{r^2 - (X - x_c)^2} \tag{11}$$

The center point and the radius r of the circle-arc are computed using the equations 12 through to 14.

$$x_c = \frac{(x_1^2 + y_1^2)*(y_3 - y_2) + (x_2^2 + y_2^2)*(y_1 - y_3) + (x_3^2 + y_3^2)*(y_2 - y_1)}{2\left[x_1*(y_3 - y_2) + x_2*(y_1 - y_3) + x_3*(y_2 - y_1)\right]} \tag{12}$$

$$y_c = \frac{(x_1^2 + y_1^2)*(x_3 - x_2) + (x_2^2 + y_2^2)*(x_1 - x_3) + (x_3^2 + y_3^2)*(x_2 - x_1)}{2\left[y_1*(x_3 - x_2) + y_2*(x_1 - x_3) + y_3*(x_2 - x_1)\right]} \tag{13}$$

$$r = \sqrt{(x_1 - x_c)^2 + (y_1 - y_c)^2} \tag{14}$$

More detailed information regarding the symmetric and simplified version can be found in Livingston and Kim (2008; 2009; 2010).

Synthetic functions

Kim et al. (2008) define the synthetic function (SF) as the weighted average of ID equating and any chosen equating method (e.g., Tucker Linear, Circle-Arc, etc.) using a pre-specified weighting system. The weight $(1-w)$ given to ID equating can range between 0 and 1. Therefore, the amount of equating error can be controlled. However, there is no universal guideline of creating SFs of equating methods, and thus, there is always a heuristic need to further investigate which weighting system works better than the other equating methods of interest under varying testing conditions. These testing conditions may include different examinee sizes, various test form difficulty levels and anchor item ratios, etc.

As aforementioned, equating test forms with small samples may lead large equating error. Kim, von Davier, and Haberman (2008) show the amount of the SEE when the SF is preferred over a regular equating method. To exemplify, assume that the SF of $e_y(x)$ is specified as follows:

$$e_{SF(y)}(x) = w*e_y(x) + (1-w)*e_{ID}(x) \tag{15}$$

Taking the variance and then the square root of both sides in equation 15 results in equation 16 as shown below:

$$SEE(e_{SF(y)}(x)) = w^{*} SEE(e_{y}(x)) \quad (16)$$

As Kim et al. (2008) indicated, the use of SF reduces the amount of the SEE one-half when the equal weighting system is given to both the ID and chosen equating methods. However, the equating bias is mostly introduced by ID equating under the assumption that the chosen equating method produces less or no equating bias compared to the ID equating with the same examinee data when equal weights are used to create the SFs (Kim et al., 2008). More detailed information about SFs can be found in Kim, von Davier, and Haberman (2008).

Purpose and Significance of Research

Those studies (Babcock et at., 2012; Harris, 1993; Kim et al., 2008) have greatly contributed to the small sample test equating literature. However, there is still an uncertainty for selecting the appropriate equating method to equate the test form using very small samples ($n < 50$) when the average test form difficulty and group ability differences are varied. The CA and the NWM equating methods perform better than the other equating methods under certain conditions when the sample size is particularly small, but including the SFs, they have not been widely investigated, yet. In addition, the use of the SF in equating also showed better performance than the unaltered versions of chosen methods (e.g., Chained Linear, Chained Mean, etc.), but the SFs of the CA and NWM equating methods, which are specifically proposed for small sample equating, have never been tested. As indicated by Kim et al. (2008), the SFs of the selected equating methods seems to have a potential to reduce the standard error of equating, but the accuracy of equating results have not been investigated in detail. Briefly, there is a heuristic need to further investigate behavior of the equating methods proposed especially for the small sample test score equating. Following research questions are addressed in this study;

Are there any significant interaction effects among the study factors that influence the equating accuracy?

Does the use of SFs of the CA and NWM equating methods (CAS and NWS) improve the accuracy of equating results compared to the ID equating, unaltered CA and NWM equating methods across all study conditions?

What conclusions and suggestions can be drawn regarding the choice of SFs of the NWM and CA (NWS and CAS) equating methods in small sample test score equating?

This study took a step forward from previous studies to fill the gap in the equating literature by testing the SFs of the identity (ID), Nominal Weights Mean (NWM), and Circle-Arc (CA) equating methods to explore if any other equating can be used to equate test forms when the difficulty levels of the test forms are different. To the best of my knowledge, no study exists in the literature that investigates the performances of the ID, NWM, and CA equating methods including the SFs under the NEAT design using very small samples. Thus, it is my belief that this study would be useful for those who need a practical guideline to equate the test forms using these equating methods under the conditions considered in this study.

Related Researches

There is a vast literature on test score equating; however, only few people in the field have been conducting research on small sample equating in recent years (see, for example, Kim & Livingston, 2011; Livingston & Lewis, 2009; Kim et al., 2006). Some researchers investigated the behaviors of equating methods mainly focusing on the size of the examinees samples (e.g. Hanson et al., 1994; Livingston, 1993; Parshall, Du Bose, Houghton, & Kromrey, 1995). They also tested the effect of the data smoothing techniques on equating accuracy. Findings from those studies showed that increasing the sample size and/or smoothing the examinee score distributions both reduced the equating error, but the larger degree of smoothing is applied to raw data, the more systematic equating error has been observed (e.g, Livingston, 1993).

An extant small sample equating studies have been conducted after year 2005 by several researchers (e.g. Babcock et al., 2012; Kim et al., 2011; Kim et al., 2006; Livingston & Kim, 2009, Skaggs, 2005). Their primary focus was to either propose new equating method or integrate the existing equating methods to overcome the sample size limitation in test score equating. Skaggs (2005) paid attention to

the to the test form difficulty differences given in standardized mean difference (SMD) units since the test form difficulty differences mainly affect the equating accuracy. Other researcher therefore also considered the SMD between the test form difficulty levels to decide whether equating was necessary or which equating method should have been used.

Kim, von Davier, and Haberman (2006), for example, compared the SF of the Chained Linear (CL) method with the ID equating under the NEAT design. They examined the effect of external and internal anchor items on the accuracy of equating results. Various sizes of examine samples (n = 10, 25, 50,100, 200) from a national assessment data were used. Their findings indicated that the use of the SF performed better when the sample size was smaller than 200 in terms of total equating error, but they recommend the use of the ID equating with examinee sample size of 25 when the test forms shows equal or very similar difficulty levels (less than .10 SMD).

Livingston and Kim (2009) compared the CE, LL, CL, CM, and the ID equating methods using a teacher certification examination data. Small samples were drawn to equate the test forms which showed substantial differences in their difficulty levels (.36 SMD). The CA method performed better than the other chosen methods especially for equating the test scores at the extreme ends on the score scale when the SMD was less than .10 and the sample size was over 150. However, they maintained that the ID was the most favorable for the test forms with .10 differences when the sample size was less than 100.

Sunnassee (2011) tested the performance of the ID, CA, CL, Tucker Linear (TL), Levine Linear (LL), presmoothed CE and Frequency Estimation (FE) methods in a simulation study. Five different sample sizes (25, 50, 100, 200, and 400) were compared to equate the test forms of which the difficulty differences were ranged from .0 to .75 SMD. The findings showed that all the equating methods were capable of adjusting the test form difficulty differences when the difficulty levels of the test forms were equal to or less than .25. However, both the CA and the ID equating methods produced large bias especially when the ability levels of the examinees differed much on average.

Babcock, Albano, and Raymond (2012) compared the NWM, smoothed EE, TL, SF of the TL method, CA, and the ID equating with the small examinee samples (n = 20, 50, and 80). The ability levels of the new form examinee group were specified as less than, equal to, and larger than the ability group of examinees taking the base test form. Test forms with different difficulty levels were also used to compare those equating methods. Their findings indicated that the ID equating was the most accurate among the others when the test forms were equal in their difficulty levels. However, the NWM method was the most favorable ones when the group ability levels were not equal to each other. Both the CA and NWM equating methods performed well when the test form difficulty levels and the group ability levels differed much.

METHOD

The central objective of this study was to compare the performance of the chosen equating functions and their synthetic equating functions under a variety of conditions. The relationship among those equating methods is given in Figure 2. A series of computer simulations were therefore carried out under a variety of testing conditions that potentially affect the performance of the equating accuracy. Several sampling factors and psychometric properties of the test forms used in the equating procedure were considered. Those factors were the sample size of the new form test takers, the SMD between the examinee groups' ability levels, and SMD between test form difficulty levels.

Sample Size

Resampling studies in the literature show that the examinee samples taking the new test form typically range from 10 to 100 in teacher certification or licensure examinations (e.g. Babcock et al., 2012; Kim et al., 2006; Livingston & Kim, 2009). In the present study, the sample sizes of the new form test takers were also specified as 10, 25, 50, and 100, but it was held constant at 200 examinees taking the base test form. Four levels of the sample sizes ($\Delta_N = 10, 25, 50, 100$) were therefore used to simulate the actual testing condition. To establish the criterion equating function, 50,000 examinees' response data were used for each of the test forms.

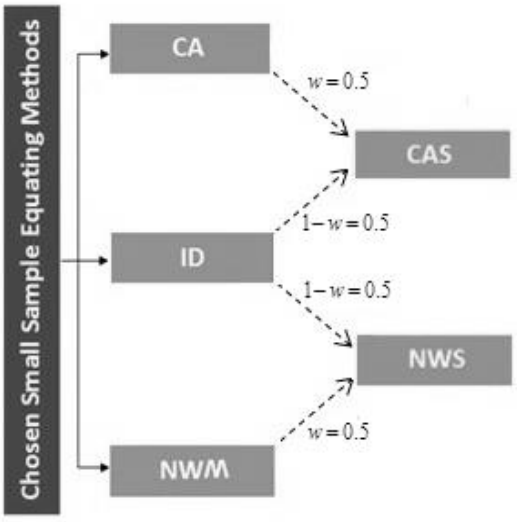


Figure 2. The Chosen Equating Methods and Associated Synthetic Functions

Group Mean Ability Levels

The group mean ability differences can be conceptualized in standardized mean differrences (SMD) units. In this study, the examinee group taking the base or the reference test form were sampled from the standard normal distribution for simplicity $[\theta_{Base} \sim N(0,1)]$. For the examinee group taking the new test form, four different group mean ability levels were chosen from the four normal distributions which have different means, but same standard deviation. In addition, symmetric examinee distributions with negative and positive means were specified to determine how lower and higher ability groups affect the equating results $[\theta_{New} \sim N(-.15,1); N(-.03,1); N(+.03,1); N(+.15,1)]$.

It is important to note here that, in test score equating, the groups' mean ability differences between $\Delta_\theta = .05$ and $\Delta_\theta = .1$ SMD are generally considered very large, and an ability difference of .25 SMD between the examinee groups is considered extremely large difference (Wang et al., 2008). According to this rule of thumb, the groups' mean ability differences (Δ_θ) were intentionally specified in SMD units as small and large differences to investigate its effect on the equating accuracy for the methods tested in this study.

Test form mean difficulty levels

Mean difficulty differences between test forms (Δ_δ) can be also examined in the SMD units. In Item Response Theory (IRT), the test form difficulty level is conceptualized by the average of the item difficulty statistics. Smaller δ (delta) value, for example, is an indication of a more difficult test than its associated base or reference form; likewise, larger δ (delta) value means that the new test form is less difficult than its base test form. For the present study, the average test form difficulty level of the new test form was specified with four different normal distributions $[\delta_{New} \sim N(-.20,1); N(-.05,1); N(+.05,1); N(+.20,1)]$ while the true values of δ parameter were randomly drawn from the standard normal distribution of $\delta_{Base} \sim N(0,1)$. According to Kim (2014), the SMD of .05 is considered a small difference between test forms in their average difficulty levels, but the SMD of .20 is an indication of large difference between the test form difficulty levels. With this rule of thumb, less and more difficult new test forms than the base test forms were created to investigate the effect of change in mean difficulty difference between the test forms on the equating accuracy for the equating methods tested in this study.

Data Generation Procedure

A few steps were employed to generate the examinees' item responses for each of the simulation conditions. A total of 64 (4x4x4) conditions were established with the given ability and test form

difficulty levels to compare the performances of five equating methods as indicated in Figure 2. As the first step, the true item parameter values were generated from the test form difficulty distributions for a 120 item test. Those true values were then used to create the examinees' item responses for each of the given examinees' ability distributions using the 3-Parameter Logistic (3-PL) IRT Model.

IRT establishes a relationship between latent variables and their manifestations using a monotonically increasing function that is specified in a mathematical form including person and item parameters to predict observed responses on a test item (de Ayala, 2009). IRT models can be also used to simulate data based on the psychometric characteristics of an item (difficulty, discrimination, and guessing parameters) and examinee's given ability (theta) level to determine the probability of answering each test item correctly. The mathematical expression of the 3-PL IRT Model is given in equation 17,

$$P(X_{ij}=1 \mid \theta_j, a_i, b_i, c_i) = c_i + (1-c_i)\frac{\exp(a_i(\theta_j - b_i))}{1+\exp(a_i(\theta_j - b_i))} \quad (17)$$

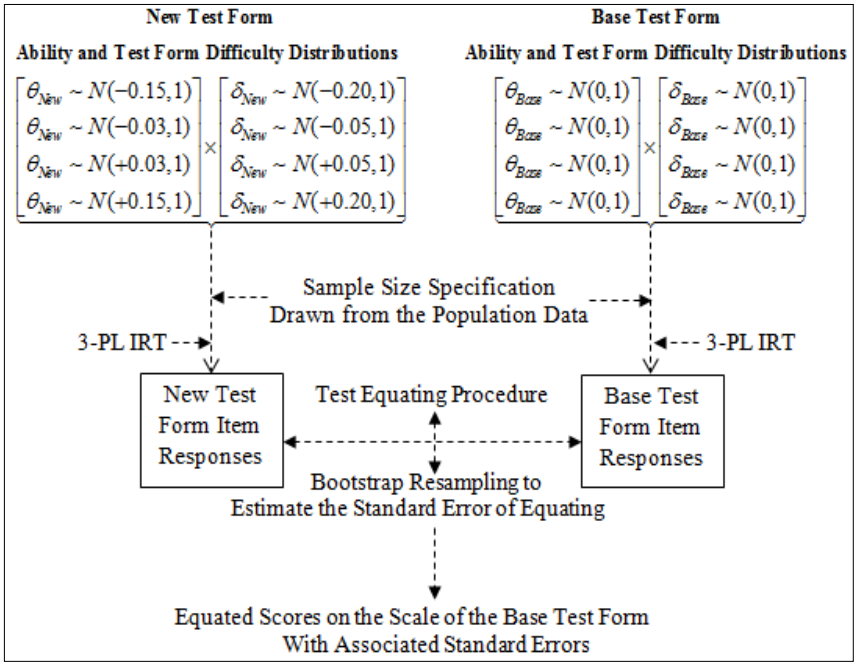


Figure 3. Data Generation and Test Equating Procedures

where the $P(.)$ is the probability of correctly answering item *i*, which is conditioned on the ability level (θ_j) of person *j* and the values of the discrimination parameter (a_i), difficulty parameter (b_i), and the guessing parameter (c_i) of item *i*. In this study, the 3-PL IRT model was used to generate examinees' binary item responses using *R* software Version 3.2.0 (R Development Core Team, 2015). R scripts were written to equate the test forms using *equate* package. Different R packages were also used in the data generation and computer simulation procedures such as *matrixStats*, *psy*, *ggplot2*, *plotrix*, and *gridExtra*.

Total number of test items is typically ranged between 80 and 120 in test forms used in small sample equating (Babcock et al., 2012; Kim et al., 2006; Livingston & Kim, 2009). Therefore, the numbers of the items in the total test and anchor test were specified as 120 and 36 respectively in order to mimic the actual test structure. Four new test forms and four base test forms were created using the true item parameter distributions. Consequently, a total number of 32 pseudo test forms responses were produced from the product of given test difficulty and ability distributions. Those examinees' responses were then used in the equating procedure as shown in Figure 3. The psychometric characteristics of test forms were given in Appendix.

Evaluation of Equating Accuracy

The accuracy of equating results under varying test conditions were evaluated by using weighted counterparts of several measures of accuracy indices including Weighted BIAS, Weighted SEE, and Weighted RMSE. To better compare the accuracy of each equating results in the full examinee population, the average of those accuracy measures were weighted using the frequency distribution of each raw score value in the new test form large group population equating. Sums of the weighted accuracy measures for each score point were then used as the overall summary measures for each of the equating methods tested in this study.

BIAS of an equated score $[\hat{e}_y(x_k)]$ is defined as the difference between the estimated average equivalent score across R replications ($[\bar{\hat{e}}_y(x_k)]$ and the true equivalent score in the population $[e_y(x_k)]$. Weighted Root Mean Squared BIAS (WBIAS) is also calculated for each score point k on the score scale to prevent the negative and positive values that cancel each other using the proportion of raw score k in the large group criterion equating. The proportion P_k provides more accurate representation of the frequency of the new test form examinees' scores in the large group population critierion equating. Considering all of the score points, the WBIAS can be formulized as follows:

$$WBIAS[\hat{e}_y(x_k)] = \sqrt{\sum_{k=0}^{K} P_k \{BIAS[\bar{\hat{e}}_y(x_k) - e_y(x_k)]\}^2} \tag{18}$$

SEE is defined as the square root of the averaged squared difference between the estimated equivalent score $[\hat{e}_y(x_{kr})]$ in the r^{th} replication and the average equivalent score $[\bar{\hat{e}}_y(x_k)]$ across R replications. Weighted SEE (WSEE) can be formulated using the proportion of raw score k in the large group criterion equating as follows:

$$WSEE[\hat{e}_y(x_k)] = \sqrt{\sum_{k=0}^{K} P_k \{\frac{1}{R}\sum_{r=1}^{R} \{\hat{e}_y(x_{kr}) - \bar{\hat{e}}_y(x_k)\}^2\}} \tag{19}$$

RMSE is defined as the squared root of the sum of the squared BIAS and the squared SEE. Weighted RMSE (WRMSE) is obtained for a score. An equated score can be obtained using the equation 20.

$$WRMSE[\hat{e}_y(x_k)] = \sqrt{(WBIAS[\hat{e}_y(x_k)])^2 + (WSEE[\hat{e}_y(x_k)])^2} \tag{20}$$

Population Criterion Equating

One way to evaluate the accuracy of equating results is to compare the equated scores obtained from the sample and the population data. In other words, a true criterion can be considered to evaluate the equating results if the equating relationship in the population is known (Harris & Crouse, 1993). Chained Linear (CL) or Chained Equipercentile (CE) equating methods are commonly used under the NEAT design to establish the equating relationship between the test forms using the population data (Livingston, Dorans, & Wright, 1990).

The equating relationship between the test forms could be unstable and inaccurate with score distributions which contain irregular score patterns (Liou & Cheng, 1995). Thus, an application of some sort of smoothing to the raw examinee data prior to equating is often recommended (Hanson, 1991; Kolen & Brennan, 2004; van der Linden & Wiberg, 2010). In this study, the CE equating with 6-univariate and 2-bivariate log-linear presmoothing were used as the crierion equating function to evaluate the accuracy of the equating results. The fit statistics (e.g., AIC, BIC, Chi-Square, etc.) of the presmoothed model were carefuly inspected to decide what extent the smoothing should be applied to the raw examinees' score data. In addition, the central tendency measures before and after smoothing were also compared in order to make sure that the smoothing did not change the shape and location of the raw score distributions.

Analysis of Equating Results

This study has four factors: (1) sample size, (2) group mean ability difference, (3) test form difficulty difference, and (4) the equating methods used to equate the test forms. Since the same data were used to investigate the performance of the equating methods in each simulation condition, the use of Mixed-Factorial ANOVA (MFA) was the most appropriate way to investigate the main and interaction effects of data generation factors with repeated-measures.

Before conducting the MFA for each of the weighted accuracy measures, the data were screened to investigate whether there was any problem in the data that may violate the ANOVA assumptions. First, no outliers were detected in the data. Standardized residuals were approximately normally distributed according to the non-significant Kolmogrov-Smirnov and Shapiro-Wilk tests. Quantile-by-Quantile (Q-Q) and Stem-Leaf plots also visually confirmed that the normality assumption for the data appeared to be satisfactory. According to the non-significant Levene's test for each of the weighted accuracy measures, the homogeniety of variance assumption was also appeared to be satisfactory for the data generation factors.

F-statistics with adjusted degress of freedom (*df*) were used to interpret the significance level of the main and interaction effects of the study factors due to significant Mauchly's test of sphericity for the equating methods or the repeated-measures. Depending on the magnitude of the correction factor (ε) reported in Mauchly's test, the interpretations of *F*-statistics were made based on either Greenhouse-Geisser's (G-G) or Huynh-Feldt's (H-F) corrections.

Two general themes were considered to present the results of this study: (1) The overall comparison of the chosen equating methods; (2) The evaluation of the main and interaction effects of the study factors. Effect size (ES) estimates (eta-squared (η_T^2)) were provided to discuss the magnitude of the main and interaction effects, instead of reporting *F*-statistics directly. The ratio of the amount of the variance explained by a main or interaction effect to the total amount of the variation was used as an effect size measure associated with each main and interaction term in the ANOVA table. An eta-squared estimate that was larger than .01 $(\eta_T^2 \geq .01)$ was considered as a threshold value to classify an effect size as an "important effect". In other words, the main or interaction effect that accounts for at least 1% of the total variation was considered as an important effect on the equating accuracy.

RESULT and DISCUSSION

Several study factors were considered to test the behavior of the ID, CA, CAS, NWM, and NWS equating methods. Table 1 shows the main and interaction effects of those factors for each measure of accuracy indices. According to the MFA results, the most important main and/or interaction effects of the study factors for the WRMSE and WBIAS measures were Δ_θ and Δ_δ, and $\Delta_E * \Delta_\theta * \Delta_\delta$. Correspondingly, Δ_N was the only between-subject factor with an important effect on the WSEE measure, but its interaction with Δ_E was the most important within-subject factor, which simply means that the chosen equating methods produce different amount of SEE when the sample size varies.

Based on the magnitudes of the effect size estimates, the interaction term of $\Delta_E * \Delta_\theta * \Delta_\delta$ explained 18% of the total variation in the WRMSE, and 28% of the total variation in the WBIAS. Eleven percent of the variation in the WSEE was accounted for by the interaction effect of $\Delta_E * \Delta_N$. This means that the total equating error (WRMSE) was affected by several factors: (a) the method used to equate the test forms, (b) the magnitude of the ability differences between the examinee groups, (c) taking the new and base test forms, (d) the magnitude of the test form difficulty differences, and (e) the size of the examinee samples taking the test. More specifically, the interaction among $\Delta_E * \Delta_\theta * \Delta_\delta$ was the main source of the systematic error in equating (WBIAS), and the interaction between Δ_E and

Table 1. Mixed-Factorial ANOVA Results

	WRMSE						***WSEE***						***WBIAS***					
Source	***SS***	***df***	***MS***	***F-ratio***	***p***	***ES***	***SS***	***df***	***MS***	***F-ratio***	***p***	***ES***	***SS***	***df***	***MS***	***F-ratio***	***p***	***ES***
Between Factors	**837.0**	**127**					**49.03**	**127**					**1010.37**	**127**				
Δ_θ	569.22	3	189.74	1082.01	.00	*.21**	.04	3	.01	1.20	.32	.00	695.80	3	231.93	1130.57	.00	*.31**
Δ_δ	5.29	3	1.76	10.06	.00	.00	.11	3	.04	3.10	.03	.00	7.45	3	2.48	12.11	.00	.00
Δ_N	14.63	3	4.88	27.81	.00	*.01**	47.46	3	15.82	1283.60	.00	*.03**	.14	3	.05	.22	.88	.00
$\Delta_\theta * \Delta_\delta$	225.98	9	25.11	143.19	.00	*.08**	.11	9	.01	1.00	.45	.00	282.81	9	31.42	153.17	.00	*.13**
$\Delta_\theta * \Delta_N$	2.44	9	.27	1.55	.15	.00	.12	9	.01	1.12	.36	.00	.93	9	.10	.50	.87	.00
$\Delta_\delta * \Delta_\delta$	1.74	9	.19	1.10	.38	.00	.09	9	.01	.85	.57	.00	1.95	9	.22	1.06	.41	.00
$\Delta_\theta * \Delta_\delta * \Delta_N$	6.51	27	.24	1.38	.15	.00	.29	27	.01	.86	.65	.00	8.16	27	.30	1.47	.10	.00
Error	11.22	64	.18				.79	64	.01			.00	13.13	64	.21			
Within Factors	**1878.91**	**398.64**					**1325.22**	**259.22**					**1235.03**	**396.70**				
Δ_E	1168.58	3.11	375.23	3731.50	.00	*.43**	1153.98	2.03	569.82	10184.02	.00	*.83**	394.81	3.10	127.39	1368.88	.00	*.18**
$\Delta_E * \Delta_\theta$	93.46	9.34	10.00	99.48	.00	*.03**	.57	6.08	.09	1.67	.13	.00	161.93	9.30	17.42	187.15	.00	*.07**
$\Delta_E * \Delta_\delta$	15.96	9.34	1.71	16.99	.00	*.01**	1.14	6.08	.19	3.37	.00	.00	22.26	9.30	2.39	25.73	.00	*.01**
$\Delta_E * \Delta_N$	65.97	9.34	7.06	70.22	.00	*.02**	156.99	6.08	25.84	461.81	.00	*.11**	.45	9.30	.05	.52	.86	.00
$\Delta_E * \Delta_\theta * \Delta_\delta$	489.40	28.03	17.46	173.64	.00	*.18**	.77	18.23	.04	.75	.75	.00	624.00	27.89	22.37	240.40	.00	*.28**
$\Delta_E * \Delta_\theta * \Delta_N$	9.55	28.03	.34	3.39	.00	.00	.98	18.23	.05	.96	.51	.00	1.07	27.89	.04	.41	1.00	.00
$\Delta_E * \Delta_\delta * \Delta_N$	2.03	28.03	.07	.72	.85	.00	.82	18.23	.04	.80	.70	.00	1.51	27.89	.05	.58	.96	.00
$\Delta_E * \Delta_\theta * \Delta_\delta * \Delta_N$	13.91	84.09	.17	1.65	.00	*.01**	2.72	54.68	.05	.89	.68	.00	10.54	83.68	.13	1.35	.05	.00
Error	20.04	199.32	.10				7.25	129.61	.06				18.46	198.35	.09			
Total	**2715.95**	**525.64**					**1374.25**	**386.22**					**2245.39**	**523.70**				

*. Important Effect (*ES* larger than 1% or $ES\ (\eta_t^2) \geq .01$).

Δ_θ : New Group Mean Ability Difference; Δ_δ : Test Form Difficulty Difference ; Δ_N : New Group Sample Size; Δ_E : Equating Methods

Δ_N was the most important factor affecting the random error in equating (WSEE). Those findings are parallel with those of Babcock et al. (2012), Livingston (1993), and Kim et al. (2006, 2008).
However, it is important for practitioners to know which equating method performs better than the others in a specific condition. Therefore, the pairwise comparisons of the equating methods are crucial to reach an overall conclusion about the performances of the chosen equating methods.
Table 5 shows the pairwise comparisons of the chosen equating methods under all simulation conditions. Some of the condition-specific comparisons of the chosen equating methods are provided in Figure 4. In addition to the WBIAS and WSEE, the WMRSE associated with each of the equating methods can be determined using the distance from the origin to the equating method of interest since both WBIAS and WSEE are the orthogonal components of the WRMSE (see, Equation 20). According to the post-hoc tests results, the ID, NW, and NWS equating methods were compatible in most of the simulation conditions. The magnitudes and directions of Δ_θ and Δ_δ played a very important role to decide which method was the most suitable to equate the test forms.
The sum of the magnitudes of Δ_θ and Δ_δ are seemed to be very helpful for deciding which equating method should be used to equate test forms with small samples. For example, when the new group examinee sample size is 20, the new the use of the NWS method is preferable to the ID equating because the sum of the Δ_δ (= .05 SMD) and Δ_θ (=.03 SMD) is equal to .02 SMD due to the same mathematical sign. In other words, if the directional shape of the examinee score distributions are same, the use of the NWS method is preferable to the ID equating when the sum of the SMDs are equal to or smaller than .08. For the conditions, the sum of the SMDs are in between 0.10 and 0.15, the use of the NWM equating is the most favorite even with the samples of 10. On the other hand, not equating is more appropriate when the sum of the SMDs is in between .20 and .35 due to the amount of total error produced by the ID equating. Those findings are comparable with those in Skaggs (2005) and Heh (2007). However, the use of the NWS now became an alternative to the ID method for equating the test forms with samples less than 50 under the NEAT design according to the findings, which has never been suggested in any other study so far.
As can be seen in Figure 4, the standard error associated with the ID equating 0 across all conditions. However, the magnitude of the systematic error or the equating bias was quite substantial. For all the equating methods, except for the ID equating, the magnitude of the standard error was reduced while the sample size increased. Similar patterns were also observed when the size of the examinee samples got larger. This result is parallel with those from Babcock et al. (2012) and Kim et al. (2008). Similar to the findings in Skaggs (2005), the magnitude of the equating bias got smaller when the magnitude of the SMD in test form difficulty levels and group ability levels got smaller ($\Delta_\delta \pm 0.05$ and $\Delta_\theta = \pm 0.03$).

CONCLUSION

One of the requirements in test score equating is to have large sample in order to obtain accurate results, but this may not be the case in real classroom settings or in teacher certification/ licensure examination programs. In the present study, I tested and compared the performance of several equating methods under varying testing conditions that may represent a real testing scenario where some of the factors that influence the equating accuracy were manipulated. More specifically, the SFs with an equal weighting system were used to form a compromise between the ID equating and the CA and NWM equating methods, respectively.
The findings show that the use of the ID equating or the SF of the NW method is preferable to the use of the unaltered version of the NW and CA even with samples less than 50, but with the test forms that are similar in their psychometric characteristics. The use of a traditional equating method with very small samples would be extremely harmful than the use of the ID equating due to the effect of the small samples on the random equating error (WSEE). However, the use of the NWS or the ID equating methods produced more accurate results in terms of the total equating error (WRMSE) for the conditions in which the difference between the test forms or the difference in the shape of their respective score distributions was not substantial.

Table 2. Pairwise Comparison of the WRMSE Estimates of the Chosen Equating Methods

Δ_N	Δ_θ	Δ_δ	Δ_M^*	Δ_θ	Δ_δ	Δ_M^*
10	-.15	-.20	NW < NWS < ID < CAS < CA	.03	-.20	ID < NWS < CAS < NW < CA
	-.15	-.05	NW < NWS < ID < CAS < CA	.03	-.05	ID < NWS < NW = CAS < CA
	-.15	.05	ID < NWS < CAS = NW < CA	.03	.05	NWS = ID < NW < CAS < CA
	-.15	.20	ID < NWS < CAS < NW < CA	.03	.20	NWS = ID < CAS < NW < CA
	-.03	-.20	NWS = ID < CAS = NW < CA	.15	-.20	ID < NWS < CAS = NW < CA
	-.03	-.05	NWS = ID < NW < CAS < CA	.15	-.05	ID < NWS < NW < CAS < CA
	-.03	.05	ID < NWS < CAS < NW < CA	.15	.05	NW < NWS < ID = CAS < CA
	-.03	.20	ID < NWS < CAS < NW < CA	.15	.20	NW < NWS < CAS = ID < CA
20	-.15	-.20	NW < NWS < CAS = CA = ID	.03	-.20	ID < NWS < CAS < NW < CA
	-.15	-.05	NW < NWS < ID = CAS < CA	.03	-.05	ID < NWS < CAS = NW < CA
	-.15	.05	ID < NWS < CAS = NW < CA	.03	.05	NWS < ID < NW < CAS < CA
	-.15	.20	ID < NWS < CAS < NW < CA	.03	.20	NWS < ID < CAS < NW < CA
	-.03	-.20	NWS < ID < CAS < NW < CA	.15	-.20	ID < NWS < CAS = NW < CA
	-.03	-.05	NWS < ID < NW < CAS < CA	.15	-.05	ID < NWS < NW < CAS < CA
	-.03	.05	ID < NWS < CAS < NW < CA	.15	.05	NW < NWS < ID = CAS < CA
	-.03	.20	ID < NWS < CAS < NW < CA	.15	.20	NW < NWS < CAS < CA < ID
50	-.15	-.20	NW < NWS < CA = CAS = ID	.03	-.20	ID < NWS = CAS < NW < CA
	-.15	-.05	NW < NWS < ID = CAS < CA	.03	-.05	ID < NWS = CAS = NW < CA
	-.15	.05	ID < NWS < CAS < NW < CA	.03	.05	NWS < CAS = ID = NW < CA
	-.15	.20	ID < NWS < CAS < NW < CA	.03	.20	NWS < ID < CAS < NW = CA
	-.03	-.20	NWS < CAS = ID < NW = CA	.15	-.20	ID < NWS < CAS = NW < CA
	-.03	-.05	NWS = ID < NW = CAS < CA	.15	-.05	ID < NWS < NW < CAS < CA
	-.03	.05	ID < NWS < CAS < NW < CA	.15	.05	NW < NWS < ID = CAS < CA
	-.03	.20	ID < NWS < CAS < NW < CA	.15	.20	NW < NWS < CA = CAS = ID
100	-.15	-.20	NW < NWS < CAS = CA = ID	.03	-.20	ID < NWS = CAS < NW = CA
	-.15	-.05	NW < NWS < ID = CAS < CA	.03	-.05	ID < NWS = CAS = NW < CA
	-.15	.05	ID < NWS < CAS = NW < CA	.03	.05	NWS < CAS = ID = NW < CA
	-.15	.20	ID < NWS < CAS < NW < CA	.03	.20	NWS < ID < CAS < NW = CA
	-.03	-.20	NWS = CAS = ID < NW = CA	.15	-.20	ID < NWS < NW < CAS < CA
	-.03	-.05	NWS < NW < CAS = ID = CA	.15	-.05	ID < NWS < NW < CAS < CA
	-.03	.05	ID < NWS < CAS < NW < CA	.15	.05	NW < NWS < ID = CAS = CA
	-.03	.20	ID < NWS < CAS < NW < CA	.15	.20	NW < NWS < CA = CAS = ID

*. The 4th and 7th columns show the equating methods produced the smallest WRMSE. " = " sign indicates no statistical difference between the equating methods; otherwise, " < " sign shows a statistical difference between the equating methods according to pairwise comparisons of the equating methods.

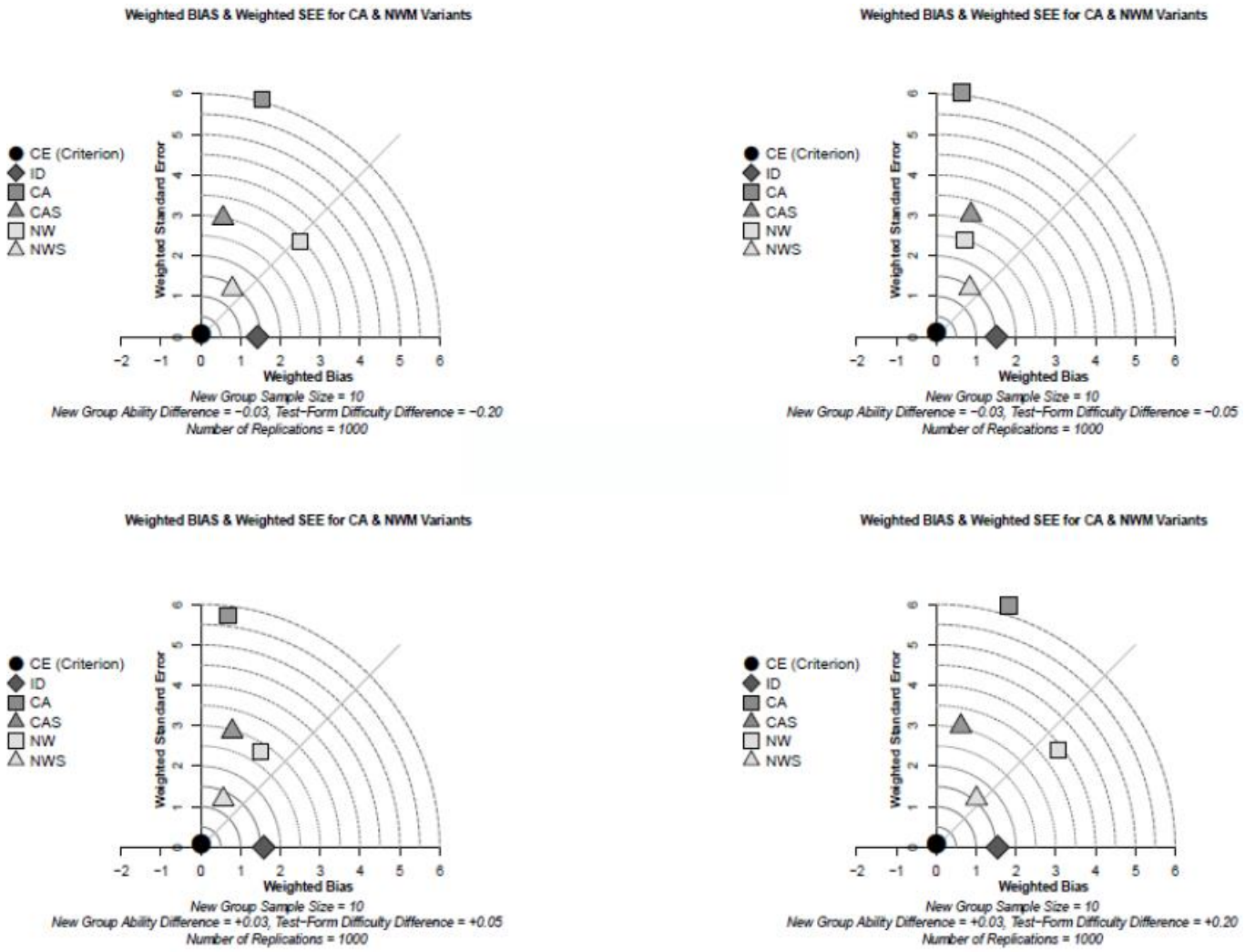
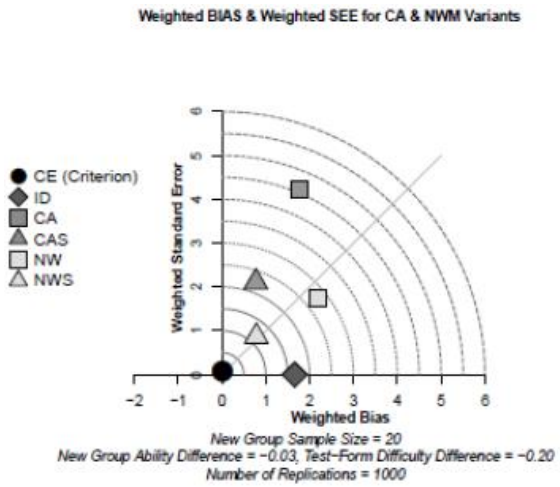
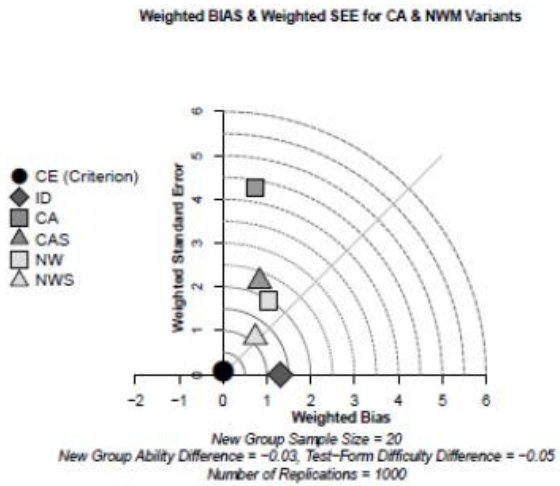
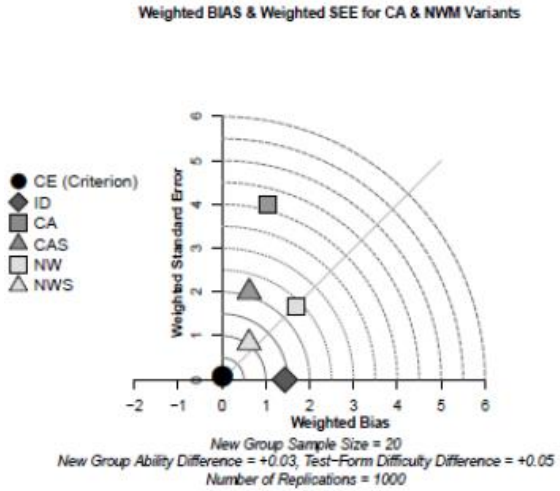


Figure 4. Some of the Condition Specific Comparisons of the Equating Methods







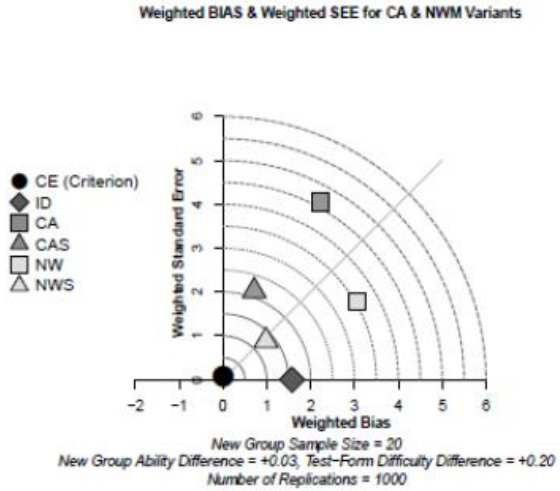


Figure 4 Cont. Some of the Condition Specific Comparisons of the Equating Methods

For example, when the magnitudes of Δ_θ and Δ_δ were small (e.g., $\Delta_\theta = \pm .03$ and $\Delta_\delta = \pm .05$) and the difference in the shape of the score distributions on the test forms were similar, the NWS produced smaller equating error (WRMSE) compared to the unaltered NWM equating method. The SF of the CA method always produced smaller equating error than its original version, regardless of the simulation condition in which they were tested. In addition, equating the test forms using the CA method also resulted with a substantial equating error (WRMSE) in all the simulation conditions. For the extreme conditions (e.g., $\Delta_\theta = \pm .15$, $\Delta_\delta = \pm .20$), the use of the ID or the NWM equating methods was more appropriate due to the amount of the total equating error.

The ANOVA results show that the systematic error (WBIAS) is mainly affected by the interaction effect of the chosen equating method, the magnitude of the group mean ability and test form difficulty differences. Correspondingly, the main source of the random equating error is the sampling fluctuations as indicated in the Table 1. As Kolen and Brennan (2004) suggest, the systematic error can be controlled with a careful test development process and also with the use of the appropriate method to equate the test forms. Even though increasing the examinee sample size would reduce the random equating error at some degree, it is very difficult for practitioners to collect more data in a certain time interval since the timely score release is one of the main goals of any examination programs after the tests administrations.

Recommendations for Practitioners

A careful test development process may eliminate the effect of test form difficulty differences on the examinee test scores to expose the examinees` ability differences. However, if the practitioners still have a concern about the test form difficulty differences after the careful test development process, then there will be a need for equating the test forms at the expense of potentially getting large equating error in small sample equating. Within the context of this study, I recommend practitioners consider two statistical measures and their mathematical signs to decide which equating method(s) or score transformation procedure(s) should be used while practicing small sample equating under the NEAT design. The first measure is the SMD on the anchor tests (see, Appendix). $SMD_{A_x - A_y}$ is used as a measure of groups` ability difference on the test since the anchor tests are taken by the two groups. The second measure is the SMD on the total test scores. $SMD_{T_x - T_y}$ is the combined measure of both group ability and test form difficulty levels to decide whether equating is necessary or which equating method should be used to equate the test forms. The SF versions of the chosen or any other equating methods can be a solution under certain conditions when the psychometric characteristics of the test forms for equating are not much different from each other. Based on the simulation conditions established in this study, Table 3 provides a basic guideline for the use of the equating method that may help while practicing test equating with small samples.

Limitations and Future Research

In this study, the computer-based simulated data, which was assumed to be normally distributed, with a limited number of factors were used. Therefore, the findings of this study should be cautiously used to make comparisons with other existing studies. Different test administration procedures may exist for each specific testing program based on the characteristics of the subject area examinations and the examinee population of interest; thus, the findings from this study should not be directly used for any specific testing program to equate the test forms using small samples. Extreme cut scores on scale score distributions were not considered in this study. Three between-subjects factors and one within-subject (as repeated measures) with a limited number of levels were investigated. The variances of the data generation factors were kept constant with a variance of 1.0 across all of simulation conditions. The SEEs of each equating method were estimated using 1000 bootstrapped samples within each simulation replication.

Further studies should be conducted using real-data from a teacher certification/ licensure examination. Also, effectiveness of the proposed equating methods should be investigated by considering different psychometric characteristics of test forms (e.g. different test lengths with varying anchor/total item ratio, internal and external anchor cases, test with low and high reliability levels, examinee groups with

varying degree of ability levels, different equating methods, mixture of different examinee score distributions, and varying sample sizes).

ACKNOWLEDGMENT

This study has been produced from author's doctoral dissertation.

The use of a meta-analysis technique in equating and its comparison with several small sample equating methods by *Caglak, Serdar*, Ph.D., The Florida State University, 2015, *Insu Paek*, Professor Directing Dissertation.

REFERENCES

Babcock, B., Albano, A., & Raymond, M. (2012). Nominal Weights Mean Equating: A method for very small samples. *Educational and Psychogical Measurement, 72*(4), 608-628.

Brennan, R. L., & Kolen, M. J. (1987). Some practical issues in equating. *Applied Psychological Measurement, 11*, 279-290.

Caglak, S. (2015). *The use of a meta-analysis technique in equating and its comparisons with several small sample equating methods* (Doctoral Dissertation). Florida State University.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: John Wiley & Sons.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

Hanson, B. A. (1991). A comparison of bivariate smoothing methods in common-item equipercentile equating. *Applied Psychological Measurement, 15*, 391-408.

Hanson, B. A., Zeng, L., & Colton, D. (1994*). A comparison of presmoothing and postsmoothing methods in equipercentile equating*. ACT Research Report 94-4. Iowa City, IA: ACT, Inc.

Harris, D. J. (1993, April). *Practical Issue in Equating*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6*(3), 195-240.

Heh, V. K. (2007). Equating accuracy using small samples in the random groups design. (Doctoral Dissertation). Retrieved from University of Ohio at https://etd.ohiolink.edu/!etd.sendfile?accession=ohiou1178299995&disposition=inline

Howell, D. C. (2007). Statistical Methods for Psychology (7th ed.). Belmont, CA: Thompson

Higher Education.

Kim, S., von Davier, A. A., & Haberman, S. (2006). An alternative to equating with small samples in the non-equivalent groups anchor tests design. *ETS Research Report Series, 2*, 1-40.

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

Kim, S. & Livingston, S. A. (2010). Comparisons among Small Sample Equating Methods in a Common-Item Design. *Journal of Educational Measurement, 47*(3), 286-298.

Kim, S., Livingston, S. A., & Lewis, C. (2011). Collateral information for equating in small samples: A preliminary investigation. *Applied Measurement in Education, 24*(4), 302-323.

Kim, H. Y. (2014). A comparison of smoothing methods for the common item non-equivalent design (Doctoral Dissertation). Retrieved from the University of Iowa at http://ir.uiowa.edu/etd/1344.

Kolen, M. J., & Brennan R. L. (2004). *Test Equating, Scaling, and Linking*. New York, NY: Springer-Verlag.

Livingston, S. A. (1993). Small sample equating with log-linear smoothing. *Journal of Educational Measurement, 30*(1), 23-39.

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

Livingston, S. A., & Kim, S. (2008). Small sample equating by the Circle-Arc method. Princeton, NJ: ETS.

Livingston, S. A., & Kim, S. (2010). Random groups equating with samples 50 to 400 test takers. *Journal of Educational Measurement, 47*(2), 15-185.

Livingston, S. A., & Kim, S. (2011). New approaches to equating with small samples, In A. von Davier (ED.), *Statistical models for test equating, scaling, and linking* (1st ed., pp.109-122).

Livingston, S. A, & Lewis, C. (2009). *Small sample equating with prior information*. Princeton, NJ: ETS.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3*(1), 73–95.

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

Peterson, N. S. (2007). Equating: Best practices and challenges to best practices. In N. J. Dorans, M. Pommerich, & P. W. Hollownd (Eds.), *Linking and aligning scores and scales* (59-71). New York, NY: Springer Science+Business Media, LLC.

R Core Team (2015). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

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

Sunnassee, D. (2011). Conditions affecting the accuracy of classical equating methods for small samples under the NEAT design: A simulation study. (Doctoral dissertation). Retrieved from the University of North Carolina at Greensboro at http://libres.uncg.edu/ir/listing.aspx?id=8164

van der Linden, W. J., & Wiberg, M. (2010). Local observed-score equating with anchor-test designs. *Applied Psychological Measurement, 34*(8), 620-640.

APPENDIXES

Appendix 1. The Psychometric Characteristics of Pseudo Test Forms

Δ_θ [1]	Δ_δ [2]	*Form* [3]	*Mean*	SD_T [4]	SD_{T_X/T_Y} [5]	SD_A [6]	SD_{A_X/A_Y} [7]	$SMD_{T_X-T_Y}$ [8]	$SMD_{A_X-A_Y}$ [9]	*Cor.* [10]	*SEM* [11]	*Rel.* [12]
-.15	-.20	X	70.12	18.77	1.00	6.03	1.00	.050	-.133	.86	4.97	.93
		Y	69.18	18.82		6.04				.86	4.61	.94
-.03	-.20	X	72.37	18.76	.99	6.04	1.00	.166	-.026	.86	4.60	.94
		Y	69.24	18.87		6.05				.86	4.62	.94
+.03	-.20	X	73.40	18.76	1.00	6.04	1.00	.227	.036	.86	4.59	.94
		Y	69.16	18.69		6.02				.86	4.94	.93
+.15	-.20	X	75.59	18.49	.98	5.97	.99	.339	.134	.86	4.53	.94
		Y	69.26	18.84		6.03				.86	4.61	.94
-.15	-.05	X	67.13	19.08	1.00	6.11	1.01	-.061	-.141	.86	4.67	.94
		Y	68.29	19.14		6.07				.86	4.69	.94
-.03	-.05	X	69.46	19.01	.99	6.06	1.00	.066	-.026	.86	4.66	.94
		Y	68.19	19.22		6.04				.86	4.71	.94
+.03	-.05	X	70.51	18.89	.98	6.08	.99	.117	.029	.86	4.63	.94
		Y	68.28	19.34		6.14				.86	4.74	.94
+.15	-.05	X	72.56	18.89	.99	6.10	1.00	.233	.131	.86	4.63	.94
		Y	68.14	19.13		6.07				.86	4.69	.94
-.15	+.05	X	65.22	18.54	.97	5.88	.99	-.295	-.133	.85	4.90	.93
		Y	70.78	19.10		5.93				.86	4.68	.94
-.03	-.05	X	67.41	18.56	.97	5.88	.99	-.181	-.024	.85	4.91	.93
		Y	70.82	19.06		5.92				.86	4.67	.94
+.03	-.05	X	68.50	18.59	.98	5.91	1.00	-.119	.032	.85	4.92	.93
		Y	70.73	19.06		5.91				.86	4.67	.94
+.15	-.05	X	70.52	18.54	.98	5.89	1.00	-.019	.129	.86	4.91	.93
		Y	70.88	18.93		5.87				.86	4.64	.94
-.15	+.20	X	62.50	19.05	1.02	6.00	.99	-.393	-.139	.85	5.04	.93
		Y	69.91	18.69		6.04				.85	4.94	.93
-.03	+.20	X	64.84	19.15	1.02	6.05	1.00	-.268	-.029	.86	5.07	.93
		Y	69.92	18.81		6.07				.86	4.98	.93
+.03	+.20	X	65.90	19.19	1.02	6.07	1.00	-.205	.030	.86	4.70	.94
		Y	69.80	18.86		6.09				.86	4.99	.93
+.15	+.20	X	68.08	19.18	1.02	6.06	1.00	-.096	.128	.86	4.70	.94
		Y	69.90	18.75		6.04				.85	4.96	.93

[1]. The new group examinee ability difference used to generate the pseudo-population data.
[2]. The new form test difficulty difference used to generate the pseudo-population data.
[3]. Form X is the new test form and Form Y is the reference (base) form in all equating procedures.
[4]. The standard deviation of Form X and Form Y test scores.
[5]. The ratio between the standard deviations of total test scores on Form X and Form Y.
[6]. The standard deviation of Form X and Form Y anchor (common) test scores.
[7]. The ratio between the standard deviations of anchor (common) test scores on Form X and Form Y.
[8]. The standardized mean difference (*SMD*) between the total test scores on Form X and Form Y.
[9]. The standardized mean difference (*SMD*) between the anchor (common) test scores on Form X and Form Y.
[10]. The correlation between the total and anchor (common) test scores.
[11]. Standard Error of Measurement.
[12]. Cronbach's Alpha reliability estimate for the total test.