International Online Journal of Educational Sciences, 2016, 8 (2), 1-11

iojes
www.iojes.net

International Online Journal of Educational Sciences

ISSN: 1309-2707

Comparison of Item Response Theory Test Equating Methods for Mixed Format Tests*

İbrahim Uysal[1], Sevilay Kilmen[2]

[1]Hacettepe University, Ankara, Turkey; [2]Abant İzzet Baysal University, Bolu, Turkey

ARTICLE INFO

Article History:
Received 30.10.2015
Received in revised form 25.12.2015
Accepted 09.01.2016
Available online 26.02.2016

ABSTRACT

This study aims to investigate the performance of test equating methods extended to mixed-format tests within the framework of Item Response Theory (IRT). To this end, a simulation study was conducted to compare equating errors of the mean/mean, mean/sigma, robust mean/sigma, Haebara, and Stocking-Lord methods under different conditions. Using 40-item tests, the effects of anchor length (10%, 20%, and 30%) and ability distribution (normal, negatively skewed, and positively skewed) were examined on a sample of 1000 participants. We used the common-item nonequivalent group design. The tests were developed using the three parameter logistic model for dichotomous simulated data and the generalized partial credit model for polytomous simulated data. The results of the study revealed that the robust mean/sigma method generally had the highest equating errors. When all conditions were evaluated, the least equating error occurred with the "Stocking-Lord" method in the case of positively skewed groups and a long anchor test (30%). Moreover, the results indicated that the groups with similar ability distributions (normal-normal, negatively skewed-negatively skewed, and positively skewed-positively skewed) produced less equation errors than the groups with different ability distributions (negatively skewed-normal, positively skewed-normal, and positively skewed-negatively skewed).

© 2016 IOJES. All rights reserved

Keywords:
test equating, item response theory, mixed format tests

Introduction

Standardized tests measure developments in students' instructional and basic skills. Standardized tests are developed by experts and are targeted for use as comprehensive objective items. Their content, instructions, implementation and scoring rules are predetermined. Measurement procedures undertaken by standardized tests address the effectiveness and quality of the test (Anastasi, 1988; Karaca, 2008; Kubiszyn and Borich, 2013; Nitko, 2004). Item types include limited/unlimited open-ended, true-false and matching items as well as multiple choice items. Item types are generally classified in two categories. The first category includes limited or unlimited open-ended items to which participants give their answers, and the second one consists of multiple choice, matching, or true-false type objective items that provide the responder with two or more options to choose from (Crocker and Algina, 1986). Objective items have consistent psychometric qualities. They are easy to respond, are effectively scored, can measure a large scope and ensure collection of evidence in a short time. Structured items have lower psychometric qualities and require higher costs for implementation and scoring, but they are more suitable to measure skills (Jodoin, 2003; Mbella, 2012).

Standardized test scores are often used to make important decisions. While some of these decisions are at the individual level (such as which university the individual will be placed or which level of the course the individual will admitted to), other decisions are at the organizational level. Decisions that shape

* This study is a part of master dissertation entitled as 'Comparison of IRT Test Equating Methods for Mixed Format Tests'.

[1] Corresponding author's address: Hacettepe University, Faculty of Education, Beytepe, Ankara, Turkey
Telephone: +903127807259
Fax: +903122992027
e-mail: ibrahim.uysal@hacettepe.edu.tr

DOI: http://dx.doi.org/10.15345/iojes.2016.02.001

© 2016 International Online Journal of Educational Sciences (IOJES) is a publication of Educational Researches and Publications Association (ERPA)

the social policy of the state are collected with these tests. Thus, it is crucial to use a structure that allows the use of different items on a common scale to ensure equality and make the right decisions (Cook and Eignor, 1991; Kolen and Brennan, 2004).

Test Equating

There are situations that require testing of the same qualities in different individuals with different tests (security, justice, etc.) (Crocker and Algina, 1986). Although test developers attempt to prepare tests as similar as possible in terms of knowledge, skills and content, it is impossible to prevent changes in test difficulty and the differentiation of content and statistical qualities from one test format to another. That is difficult for test developers, assessment experts and participants.

Test equating has been studied for 30 years and offers comparable scores when more than one test form is used (Dorans, 1990; Mohandas, 1996; Skaggs, 1990; Woldbeck, 1998). According to Kolen and Brennan (2004) and Kim and Hanson (2002), test equating is the use of different test forms interchangeably by arranging the scores in the test forms. Von Davier and Wilson (2007) define test equating process as obtaining scores that are convertible among different test formats.

Test equating procedures are undertaken based on CTT (Classical Test Theory) and IRT. However, when CTT is used, it is impossible to provide the equality and invariance assumptions (Kolen, 1981). When IRT is used, individual abilities can be identified independently of the items. This indicates whether tests are difficult or easy, and it does not affect the individuals' abilities.

One-dimensional IRT models can be used when they satisfy unidimensionality and local independence. One-dimensional tests measure a single latent ability. Local independence is related to the independence of answers provided to the items by the participants based on their abilities. IRT is generally based on the possibility of answering one item correctly. The possibility addressed here is a function generated by the individual's ability and the item difficulty. This item response function can be represented in graphic form, and it is called the test characteristic curve (Hambleton, Swaminathan and Rogers, 1991; Kolen and Brennan, 2004; Tian, 2011).

There is no need to equate IRT operations because IRT has an invariance characteristic. When calibration operations are conducted on item parameters, the abilities can be directly identified. However, this is sometimes not possible in real applications (Tian, 2011). For instance, when anchor item design is used in nonequivalent groups, samples are from different populations, and item parameters are obtained from two different tests. This situation calls for separate calibrations. When separate calibrations are implemented, it is necessary to place item parameters in a common scale. Therefore, two values are calculated to obtain a linear transformation by using equivalent methods. The *A* slope and *B* constant values are obtained via several methods that help identify the individual's ability levels in different test forms. The equation used for this calculation is provided below:

θ: ability

θ*: individual's ability level in the test that has not been taken

A: slope

B: intercept;

$$\theta^* = A\theta + B$$

The tests with anchor item design in nonequivalent groups should be as similar as possible in terms of content and difficulty. There are differences in ability among groups because the groups are assigned randomly. When this design is used, tests do not have to be implemented in the same period differences in ability distribution and can be measured based on the results obtained from common tests. The common tests can be used as criteria for the tests that are equated. Anchor items are used to identify group performances when equating the tests. The scores in one of the forms are equated to the other test form. Common tests are shorter than the regular tests, and their reliability is lower but they test the same structures. Anchor item design is a more effective design because it allows the use of non-equivalent groups and because the groups do not have to take two test forms (single implementation). However, this design

has some limitations. Group 1 that takes the A form will never take the B form; group 2 that takes the B test can never take the A test. Therefore, different models are used for equating (Angoff, 1971; Dorans, Moses and Eignor, 2010; Holland, Von Davier, Sinharay and Han, 2006; Kang and Petersen, 2009; Kolen, 1988; Petersen, Kolen and Hoover, 1993; Sinharay and Holland, 2007). Detailed information regarding these models is provided below:

Separate Calibration Methods. Separate calibration methods can be examined under two headings: moment methods (mean-mean, mean-sigma and robust mean-sigma) and characteristic curve methods (Stocking Lord and Haebara). Below you can find detailed information about these methods:

Mean-Mean. Mean-mean method defined by Loyd and Hoover (1980) calculates A and B coefficients by using the means of discrimination and difficulty parameters. The equation that will be used in the calculation is provided below:

μ(ai): mean of discrimination parameters for i scale

μ(bi): mean of difficulty parameters for i scale

μ(aj): mean of discrimination parameters for j scale

μ(bj): mean of difficulty parameters for j scale

A: slope

B: intercept

$$A = \frac{\mu(\alpha_i)}{\mu(\alpha_j)}$$

$$B = \mu(b_j) - A\mu(b_i)$$

Mean-sigma. Mean-sigma method defined by Marco (1977) uses the mean and standard deviation values of the difficulty parameter to determine A and B coefficients. The equation used in the calculation is provided below:

μ(bi): mean of difficulty parameters for i scale

μ(bj): mean of difficulty parameters for j scale

σ(bi): standard deviation of difficulty parameters for i scale

σ(bj): standard deviation of difficulty parameters for j scale

A: slope

B: intercept

$$A = \frac{\sigma(b_j)}{\sigma(b_i)}$$

$$B = \mu(b_j) - A\mu(b_i)$$

Robust mean-sigma. In contrast to the mean-sigma method, this model developed by Linn, Levine, Hastings and Wardrop (1981) addresses standard errors. The variance intensity between anchor items determine the weights in this model. First, weights for each anchor pair are calculated in the model, and these weights are scaled and weighted estimates are calculated. After identifying the means and standard deviation values of the weighted items, α and β values are determined. The α and β values include extreme data in calculating the means and the standard deviation. The α and β values do not change in each repetition (Hambleton, Swaminathan and Rogers, 1991). The equation that will be used in the calculation is provided below:

bic: difficulty parameter for i scale

bjc: difficulty parameters for j scale (basic test)

$$bic = \alpha\, bjc + \beta$$

Stocking Lord and Haebara. SL and HA methods are characteristic curve methods that take both difficulty and discrimination parameters into consideration (Hambleton, Swaminathan and Rogers, 1991). In this model, transformation constants are first identified. Then the differences between the tests or item characteristic curves are calculated. The decreases in these differences are investigated.

"Haebara Approach" was developed by Haebara (1980) and calculates the difference between item characteristic curves by taking the sums of the square of the differences in each item characteristic curve. Mathematical expression of this method is presented below:

A : Curve of the equating equation

B : Constant of the equating equation

Pij (θji, aji, bji, cji): Item characteristic function

Pij (θji, $\frac{aji}{A}$, Abji+B, cji): Equated item characteristic function.

$$Hdiff(\theta_i) = \sum_{j:V} \left[p_{i_j}(\theta_{j_i}; a_{j_j}, b_{j_j}, c_{j_j}) - p_{i_j}(\theta_{j_i}; \frac{a_{i_j}}{A}, Ab_{i_j} + B, c_{i_j}) \right]^2$$

$$Hcrit = \sum_i Hdiff(\theta_i)$$

The Stocking Lord (1983) approach calculates the difference between item characteristic curves by taking the square of the sums of differences in each item characteristic curve. Mathematical expression of this method is presented below:

A : Curve of the equating equation

B : Constant of the equating equation

Pij (θji, aji, bji, cji): Item characteristic function

Pij (θji, Abji+B, cji): Equated item characteristic function

$$SLdiff = \left[\sum_{j:V} p_{i_j}(\theta_{j_i}, a_{j_j}, b_{j_j}, c_{j_j}) - p_{i_j}(\theta_{j_i}; \frac{a_{i_j}}{A}, Ab_{i_j} + B, c_{i_j}) \right]^2$$

$$SLcrit = \sum_i SLdiff(\theta_i)$$

Bastari (2000) examined the effect of anchor item ratio and ability distribution on test equating via simulative data and mixed models. The abundance of anchor item number and similarity in distribution provides more effective outcomes. Kim and Lee (2004) compared MM, MS, HA and SL methods in mixed models based on IRT. This study examined anchor item length and ability distributions. It found that characteristic curve methods usually produce more accurate results than moment methods. Similar ability distributions produce more accurate results versus different ability distributions. Tian (2011) compared Stocking Lord and synchronic calibration methods with anchor item design in nonequivalent groups. The study utilized simulative data and showed that increases in the length of the item decreased the rate of errors. Accordingly, the IRT-based test equating methods (MM, MS, RMS, SL, HB) result in less equating error in tests in mixed models that examine the effects of anchor item ratio and ability distribution based on anchor item design in nonequivalent groups.

Method

Equating Design

Comparison of equating methods under specific factors is often used in simulation studies because real test situations are more complex and include higher number of factors (Harris and Crouse, 1993). When

simulation studies are designed with careful consideration, it is possible for them to represent real data. Researchers benefit from existing test parameters while generating simulative data (Harris and Crouse, 1993; Holland, von Davier, Sinharay and Han, 2006). Therefore, simulative studies provide a pre-model for real test data. This study utilized simulative data to contribute to the theoretical field and to identify the effects of anchor item ratio and ability distribution on equating studies on mixed models. A literature review undertaken for this study has not shown any simulative studies regarding the effect of various ability distributions.

Many designs can be used in test equating studies. The design of this study was established as anchor item design for nonequivalent groups. Kolen (1988) reported that anchor item design is commonly used in nonequivalent groups for safety and practicality.

Simulation Factors

IRT models are classified into two as dichotomously scored items and polytomously scored items. Dichotomous and polytomous items can be used separately or they can be used in the same study. This study assumed that dichotomously scored items consisted of multiple choice items, and polytomously scored items included open-ended items. 3PLM and GPCM (generalized partial credit model) were used for dichotomously and polytomously scored items, respectively. The 3PLM was selected based on the probability of providing correct answers by chance—a situation that is observed in large scale tests. In their study on mixed models to determine model-data fit by using real data (Iowa Tests of Basic Skills); Chon, Lee and Ansley (2007) reported that use of GPCM and 3PLM or 2PLM together provided the best fit. Therefore, GPCM was selected for polytomous items.

The first factor selected in the study is the anchor item ratio. Anchor items should be reliable and at sufficient length to reveal the data. Only in this way can a good fit can be obtained. According to Kolen & Brennan (2004), the rate of anchor items in mixed models with 40 or more items should be 20% or higher. In this study, the anchor item ratio was 10%, 20% and 30%, and the effect of anchor item ratios on test equating was examined. Anchor items can be selected from dichotomously or polytomously scored items or both. In this study, anchor items were selected from both dichotomously and polytomously scored items. Tate's (2000) study proposed the use of mixed anchor items because anchor items in mixed models represent the test better in terms of content and qualities. Tian's (2011) study also reported the item rate that is generally scored in a multiple manner and changes between 20% and 40% of the entire test. Table 1 presents the number of dichotomous and polytomous items in tests that may or may not be common.

Table 1. Different types of item numbers in tests

Form	Total Items	Dichotomous Items	Polytomous Items	Unique Dichotomous Items	Unique Polytomous Items	Anchor Items	Dichotomous Anchor Items	Polytomous Anchor Items
A B	40	30	10	27	9	4	3	1
C D				24	8	8	6	2
E F				21	7	12	9	3

The second factor was ability distribution. According to Kolen (1985), skewed distribution can often be observed in some large-scale tests. The assessment of skewed distributions via normal distribution assumptions may generate inaccurate results. Kang and Petersen (2009) reported that skewed ability distributions should be studied. In this study, normal, positively skewed and negatively skewed distributions were compared with similar ability distributions in this study. Next, positively and negatively skewed distributions were compared with each other as well as normal distributions. In this way, the effects of 6 different ability distributions on test equating were studied. Bastari's (2000) also focused on similar and different ability distributions. Bastari reported that similar ability distributions based on N(0-1) - N(0-1) and N(0-1) - N(1-1) distributions generated less erroneous results in test equating. Figure 1 represents the ability distributions of the equated groups.

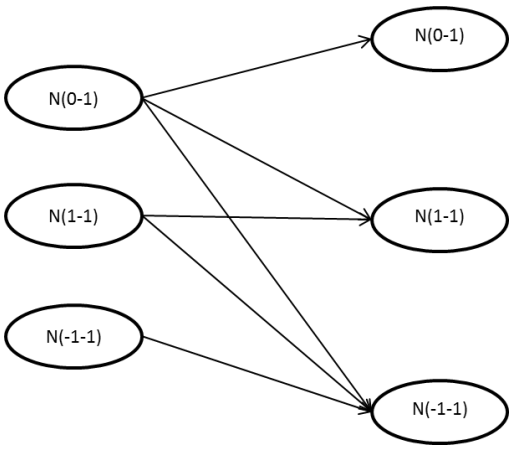


Figure 1. Ability distributions of equated groups

The study examined 900 replications—100 each for 9 structures based on ability distributions (normal, positively skewed, and negatively skewed) and anchor item ratios (4,8, and 12 items) for dichotomous and polytomous items. Han (2008) also generated 100 replications for each case. Harwell, Stone, Hsu and Kirisci (1996) stated that at least 25 replications are required in studies where parameter values are manipulated. In total, 90 conditions were investigated in the study when sub-dimensions of the variables were considered (6x5x3). These conditions are 6 cases related to ability distribution, 5 cases related to equating methods and 3 cases related to anchor item ratios.

Data Generation

This study took the sample size (1000 individuals) and test length (40 items) as constant factors. A sample of 1000 is a convenient number for getting accurate results and is an appropriate foundation for comparison (Fitzpatrick and Yen, 2001). Spence (1996) also reported that test length should be at least 35 items for a reliable calculations.

The study first generated the item parameters for tests with different anchor item ratios (4/40, 8/40 and 12/40) with the help of WINGEN3 program (Han and Hambleton, 2007). Normal distributions were used while distributions were generated. Ability distributions were generated by holding standard deviations constant and by changing the means. The means were 1.0, 0.0 and -1.0. Bastari (2000), Cao (2008), and Kim and Lee (2004) also took standard deviations as a constant and changed the means to generate distributions.

Parameters that were similar to real data were used. Discrimination, difficulty and chance parameters were respectively determined as U(0.5,1.5), b N(0,1) and c BETA(8,32) in the 3PLM. A uniform (regular) distribution was used while identifying the discrimination parameters and the 0.5-1.5 range was used for limitations to represent real data. In a study on item transformation and classification in mixed models, Montgomery (2012) used a uniform distribution for discrimination parameters. Difficulty parameters were derived from normal distribution and the mean was determined to be 0, and the standard deviation was 1. The α and β values were 8 and 32, respectively, while determining the chance parameters and beta distribution. The study assumed that polytomous items were open-ended items and scored in 5 categories. The items in the research consisted of 5 categories and were scored as 0, 1, 2, 3, and 4. A parameter for GPCM was determined to be the same with 3PLM and the b parameter was N(-1.5,0.2), N(-0.5, 0.2), N(0.5, 0.2), and N(1.5,0.2).

The psychometric qualities (difficulty and discrimination values) of the tests generated in the study were similar, and thus they were equated with a horizontal equation. Six test forms for three different anchor items were generated. This resulted in three item groups while equating the two test forms. These include anchor items (A), basic test items test (T) and items related to (K) for comparison.

Table 2. Item groups used in anchor item design in nonequivalent groups

Sample	(T)	(K)	Anchor Items (A)
Groups 1	✓		✓
Groups 2		✓	✓

This ensured that all item groups in the study had the same psychometric qualities and these test forms were placed at the end of the anchor item set tests to equate two by two. Item parameters of anchor item sets were defined to be the same in different test forms. For example, the last 4 items in test forms with 4 common items in a 40-item test have the same parameters. This is done to place ability parameters estimated via anchor items in a common scale and to form links between the test forms to be equated. In this manner, A (curve) and B (constant) values are used in the equation to ensure transformation of test scores from different test forms. Ability distributions were generated as secondary. Ability distributions were formed as normal N(0,1), negatively skewed N(1,1) and positively skewed N(-1,1). A main group and 100 comparison replicates were generated to ensure stability. The main group took the TO test form whereas the comparison groups took the KO test form. TO and KO test forms were supposed to measure the same structure.

Data Analysis

Item and ability parameters related to study data were predicted with the help of PARSCALE 4.1 program. The PARSCALE program uses a likelihood ratio chi-square method and a separate calibration to predict item parameters. Parameter estimations were undertaken for each of the TO and KO tests and each of the different ability distributions in the same calibration method (Bastari, 2000). In this study, item parameter estimates were identified for 100 groups that took the TO test after identifying the item parameter estimates for the group that took the KO test. Item parameters were placed on the same scale through calibration and item parameters obtained after transformation and the abilities of the responders were predicted again by using logistic functions. Therefore, ability distributions for each test were similar. The study utilized the EAP (Bock and Mislevy, 1982) method, which is used for universal parameters. Ability predictions were obtained by using the mean ability of the group.

Equation coefficients for different methods were identified by equating item parameters through the IRTEQ program. Therefore, the participants who took the TO form also took the KO form, and their abilities were determined through equation coefficients. The difference between abilities obtained from the TO form and the predicted abilities obtained from KO form was used to identify errors. The RMSD (Root Mean Square Differences) measure was used for this purpose. The RMSD equating criteria is a measure that is often used.

Equating Criteria. RMSD was used as the equating criteria to identify the best equating method. Mathematical description of RMSD is presented below:

θ: Real ability level

θ^*: Predicted ability level

f: Frequency for ability levels

$$RMSD = \sqrt{\frac{\sum_i f_i(\theta^* - \theta)^2}{\sum_i f_i}}$$

Low RMSD means point to the fact that the method undertakes equations with fewer errors. While RMSD means allow investigation of method accuracy, they also compare mean errors with methods (Tian, 2011). The ability parameters should be similar to obtain low RMSD means.

Findings and Discussion

Table 3 presents the findings for equating errors based on 10%, 20% and 30% anchor item ratios when the groups are normal-normal, negatively skewed-normal, positively skewed-normal, negatively skewed-negatively skewed, positively skewed-positively skewed, and positively skewed-negatively skewed. The investigation of all conditions shows that the SL method provided the least number of errors when positively skewed-positively skewed distributions were equated. Examination of error ratios generated by MM, MS, RMS, HA and SL methods based on changing anchor item ratio shows that 10% anchor item ratio provided the most errors but mean errors decreased when 20% and 30% anchor item ratio was used. This might be related to representativeness of anchor items. Bastari (2000), Hills, Subhiyah and Hirsch (1988) and Tate (2000) reported that increasing anchor item ratio from 10% to 20% significantly decreased equating errors. Meng (2012) and Cohen and Kim (1998) found that increasing the anchor item numbers decreased

equating errors. These results are similar to our findings. He's (2011) study identified that increases in anchor item ratios did not significantly affect the results. This result is contradictory to the findings of this study and the difference is believed to have been caused by the methods.

Table 3. RMSD means for different equating methods based on changing anchor item ratios and ability distributions

Ability Distribution	Sample Size	Test Length	Anchor Item Ratio	3PLM and GPCM				
				MM	MS	RMS	HA	SL
G1: N(0,1) G2: N(0,1)	1000	40	10%	0,120	0,130	0,204	0,152	0,136
			20%	0,072	0,069	0,090	0,057	0,057
			30%	0,074	0,079	0,113	0,069	0,064
G1: N(1,1) G2: N(0,1)	1000	40	10%	0,889	0,889	1,899	0,937	0,891
			20%	0,939	0,936	1,998	0,934	0,929
			30%	0,944	0,946	1,993	0,925	0,924
G1: N(-1,1) G2: N(0,1)	1000	40	10%	1,030	1,030	2,080	1,014	1,030
			20%	1,000	1,004	2,150	1,011	1,000
			30%	0,992	0,991	2,137	1,022	1,014
G1: N(1,1) G2: N(1,1)	1000	40	10%	0,104	0,097	0,174	0,126	0,099
			20%	0,060	0,109	0,102	0,053	0,055
			30%	0,126	0,080	0,060	0,061	0,069
G1: N(-1,1) G2: N(-1,1)	1000	40	10%	0,152	0,243	0,220	0,091	0,081
			20%	0,109	0,070	0,076	0,061	0,067
			30%	0,083	0,061	0,112	0,055	0,051
G1: N(-1,1) G2: N(1,1)	1000	40	10%	1,899	1,844	4,208	1,995	1,943
			20%	1,886	2,001	4,223	2,005	1,969
			30%	1,815	1,885	4,170	1,911	1,901

The RMS method had the highest error rate. In their study that compared RMS and the characteristic curve method SL, Stocking and Lord (1983) found that the SL method was superior to the RMS method. In their study which compared MM, MS, RMS and SL methods, Cohen and Kim (1998) showed that the most erroneous results were provided with the RMS method similar to our data.

The results do not suggest a single method with the lowest error rate. However, characteristic curve methods starting from 20% anchor item ratio generally result in less error. Cohen & Kim (1998), Kim and Kolen (2006) and Kim and Lee (2004) reported that characteristic curve methods provide better results than moment methods. Our results agree with the literature. Lower error ratios in characteristic curve methods can be explained via the sensitivity of these methods to the changes in the conditions. Indeed, Baker and Al Karni (1991), Cohen and Kim (1998) and Stocking and Lord (1983) showed that characteristic curve methods have structures that allow removal of inconsistencies and are more sensitive to test qualities.

Baker and Al Karni (1991) studied both horizontal and vertical equating and found that the MM method generated more accurate results than the SL method. This finding is contradictory to the findings of the current study, and the difference may be related to vertical equating practice. This study shows that the SL method generates fewer errors than the HA method. Lee and Ban (2010) found that the Haebara method generated fewer errors than the Stocking Lord method. This finding is contradictory to our findings. The difference might be related to the selected design. Kilmen (2010) reported that the SL method generated fewer erroneous results than the HA method.

Examination of error ratios shows that mean errors related to similar ability distributions are close and significantly lower than mean errors related to different ability distributions. Cao (2008), Kilmen (2010), Meng (2012) and Paek and Young (2005) found that changes in ability distribution are effective at equating. The results obtained from equating groups with similar ability distributions are more accurate than the

results obtained from equating groups with different ability distributions. These findings are consistent with the results of the current study.

Based on the test findings, we suggest using equating among groups with similar ability distributions and 20% and higher anchor item ratios in equations implemented on groups with similar ability distributions. The robust means-sigma method is not suggested for use because it generates high error rates under conditions when distributions are not similar. This research is a theoretical simulative study undertaken to identify the effects of specific factors. Similar studies can be implemented on real data. Ability distributions can be studied with different factors. Comparisons can be undertaken by using more than one criteria rather than a single equating criterion.

References

Anastasi, A. (1988). *Pschological testing*. New York: Macmillian.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Eds.), *Educational measurement* (pp. 508-600). Washington: American Council on Education.

Baker, F. B. & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement*, *28*(2), 147-162. Retrieved on November, 13, 2013 from http://www.jstor.org/stable/1434796.

Bastari, B. (2000). *Linking multiple-choice and constructed-response items to a common proficiency scale.* Unpublished doctoral dissertation, University of Massachusetts.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431-444. doi: 10.1177/014662168200600405

Cao, Y. (2008). *Mixed-format test equating: Effects of test dimensionality and common-item sets*. Unpublished doctoral dissertation, University of Maryland.

Chon, K. H., Lee, W. C. & Ansley, T. N. (2007). *Assessing IRT model-data fit for mixed format tests*. Casma Research Report, 26, November, 2007.

Cohen, A. S. & Kim, S. (1998). An investigation of linking methods under the Graded Response Model. *Applied Measurement in Education*, *22*(2), 116-130. doi: 10.1177/01466216980222002

Cook, L. L. & Eignor, D. R. (1991). IRT equating methods. *Educational Measurement: Issues and Practice, 10*(3), 37–45. doi: 10.1111/j.17453992.1991.tb00207.x

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. USA: Harcourt Brace Jovanovich College.

Dorans, N. J. (1990). Equating methods and sampling designs. *Applied Measurement in Education*, *3*(1), 3-17. doi:10.1207/s15324818ame0301_2

Dorans, N. J., Moses, T. P. & Eignor, D. R. (2010). Principles and practices of test score equating. *ETS Research Report*, 41.

Fitzpatrick, A. R. & Yen, W. M. (2001). The effects of test length and sample size on the reliability and equating of tests composed of constructed-response items. *Applied Measurement in Education*, *14*(1), 31-57. doi:10.1207/S15324818AME1401_04

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*(3), 144-149.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. USA: Sage.

Han, K. T. (2008). *Impact of item parameter drift on test equating and proficiency estimates*. Unpublished doctoral dissertation, University of Massachusetts Amherst.

Han, K. T. & Hambleton, R. K. (2007). *User's manual: WinGen* (Center for Educational Assessment Report No. 642). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Harris, D. J. & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education*, *6*(3), 195-240. doi:10.1207/s15324818ame0603_3

Harwell, M., Stone, C. A., Hsu, T. C. & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*(2), 101-125. doi: 10.1177/014662169602000201

He, Y. (2011). *Evaluating equating properties for mixed-format tests*. Unpublished doctoral dissertation, University of Iowa.

Hills, J. R., Subhiyah, R. G. & Hirsch, T. M. (1988). Equating minimum-competency tests: Comparisons of methods. *Journal of Educational Measurement*, *25*(3), 221-231. Retrieved from http://www.jstor.org/stable/1434501 on November, 12, 2013.

Holland, P. W., von Davier, A. A., Sinharay, S. & Han, N. (2006). *Testing the untestable assumptions of the chain and poststratification equating methods for the NEAT design.* Research Report, June, Educational Testing Service, Princeton, NJ.

Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement*, *40*(1), 1-15. doi: 10.1111/j.1745-3984.2003.tb01093.x

Kang, T. & Petersen, N. S. (2009). *Linking item parameters to a base scale*. Paper presented at the National Council on Measurement in Education (San Diego, California, April, 2009).

Karaca, E. (2008). Ölçme ve değerlendirmede temel kavramlar. In S. Erkan & M. Gömleksiz (Eds.), *Eğitimde ölçme ve değerlendirme* (pp. 1-36). Ankara: Nobel.

Kilmen, S. (2010). *Madde Tepki Kuramı'na dayalı test eşitleme yöntemlerinden kestirilen eşitleme hatalarının örneklem büyüklüğü ve yetenek dağılımına göre karşılaştırılması.* Unpublished doctoral dissertation, Ankara University.

Kim, S. & Hanson, B. A. (2002). Test equating under the multiple-choice model. *Applied Psychological Measurement*, *26*(3), 255-270. doi: 10.1177/0146621602026003002

Kim, S. & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education*, *19*(4), 357-381. doi:10.1207/s15324818ame1904_7

Kim, S. & Lee, W. (2004). *IRT scale linking methods for mixed-format tests* (ACT Research Report 2004-5). Iowa City, IA: Act, Inc.

Kolen, M. J. (1981). Comparison of traditional and Item Response Theory methods for equating tests. *Journal of Educational Measurement*, *18*(1), 1-11. doi: 10.1111/j.1745-3984.1981.tb00838.x

Kolen, M. J. (1985). Standard errors of Tucker equating. *Applied Psychological Measurement, 9*(2), 209-223. doi: 10.1177/014662168500900209.

Kolen, M. J. (1988). An NCME instructional module on traditional equating methodology. *Instructional Topics in Educational Measurement*, Winter, 1988.

Kolen, M. J. & Brennan, R. L. (2004). *Test equating, scalling and linking* (Second Edition). USA: Springer.

Kubiszyn, T. & Borich, G. D. (2013). *Educational testing and measurement: Classroom application and practice,* New Jersey: Wiley.

Linn, R. L., Levine, M. V., Hastings, C. N. & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Pscychological Measurement, 5*(2), 159-173. doi: 10.1177/014662168100500202

Loyd, B. H. & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*(3), 179-193. Retrieved from http://www.jstor.org/stable/1434833 on November, 12, 2013.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*(2), 139-160. doi: 10.1111/j.1745-3984.1977.tb00033.x

Mbella, K. K. (2012). *Data collection design for equivalent groups equating: Using a matrix stratification framework for mixed-format assessment*. Unpublished doctoral dissertation, The University of North Carolina.

Meng, Y. (2012). *Comparison of Kernel equating and item response theory equating methods*. Unpublished doctoral dissertation, University of Massachusetts.

Mohandas, R. (1996). *Test equating, problems and solutions: Equating English test forms for the Indonesian junior secondary school final examination administered in 1994*. Unpublished doctoral dissertation, Flinders University of South Australia.

Montgomery, M. (2012). *Investigation of IRT parameter recovery and classification accuracy in mixed format.* Paper Presented at the Annual Meeting of the Nation Council of Measurement in Education (University of Kansas, April, 2012).

Nitko, A. J. (2004). *Educational assessment of students,* New Jersey: Pearson.

Paek, I. & Young, M. J. (2005). Investigation of student growth recovery in a fixed-item linking procedure with a fixed-person prior distribution for mixed-format test data. *Applied Measurement in Education*, *18*(2), 199-215. doi: 10.1207/s15324818ame1802_4

Petersen, N. S., Kolen, M. J. & Hoover, H. D. (1993). Scaling, norming and equating. In Linn, R. L. (Eds.), *Educational measurement* (pp. 221-262). USA: The Oryx.

Sinharay, S. & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed. *Journal of Educational Measurement*, *44*(3), 249-275. doi: 10.1111/j.1745-3984.2007.00037.x

Skaggs, G. (1990). To match or not to match samples on ability for equating: A discussion of five articles. *Applied Measurement in Education*, *3*(1), 105-113. doi: 10.1207/s15324818ame0301_8

Spence, P. D. (1996). *The effect of multidimensionality on unidimensional equating with item response theory*. Unpublished doctoral dissertation, University of Florida America.

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201-210. doi: 10.1177/014662168300700208

Tate, R. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement*, *37*(4), 329-346. doi: 10.1111/j.1745-3984.2000.tb01090.x.

Tian, F. (2011). *A comparison of equating/linking using the Stocking-Lord method and concurrent calibration with mixed-format tests in the non-equivalent groups common-item design under IRT*. Unpublished doctoral dissertation, Boston College.

Von Davier, A. A. & Wilson, C. (2007). IRT true-score test equating: A guide through assumptions and applications. *Educational and Psychological Measurement*, 67(6), 940-957. doi: 10.1177/0013164407301543

Woldbeck, T. (1998). *Basic concepts in modern methods of test equating*. Paper presented at the annual meeting of the Southwest Psychological Association (New Orleans, April 11, 1998).