Int. J. Asst. Tools in Educ., Vol. 5, Issue 1, (2018) pp. 50-57

http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

Research Article

Investigation of Equating Error in Tests with Differential Item Functioning

Meltem Yurtçu[*1], **Cem Oktay Güzeller**[2]

[1]Hacettepe University, Faculty of Education, Department of Measurement and Evaluation in Education, Turkey
[2]Akdeniz University, Faculty of Tourism, Turkey

Abstract: In this study purposes to indicate the effect of the number of DIF items and the distribution of DIF items in these forms, which be equalized on equating error. Mean-mean, mean-standard deviation, Haebara and Stocking-Lord Methods used in common item design equal groups as equalization methods. The study included six different simulation conditions. The conditions were compared according to the number of DIF items and the distribution of DIF items on tests. The results illustrated that adding DIF items to tests were equated caused an increase in the errors obtained by equating methods. We may state that the change in errors is lowest in characteristic curve transformation methods, largest in moment methods depending on the situations in these conditions.

ARTICLE HISTORY

Received: 27 May 2017
Revised: 13 June 2017
Accepted: 15 September 2017

KEYWORDS

Equation error,
DIF,
IRT equation methods,
Large Scale Exams

1. INTRODUCTION

Countries participate in large-scale tests at international or national level or prepare and implement large-scale examinations in order to evaluate the educational systems or to place students in upper level educational institutions. These implemented tests are prepared in various forms in order to ensure reliability and to be able to compare the test scores of individuals taking these tests at different times. It is necessary to equate their scores in order to be able to make a comparison of scores of people taking these test forms or to make a comparison of the difficulty of exams prepared for the same purpose (Dorans & Holland, 2000; Dorans, 2004; Kim, Walker & McHale, 2010).

Through procedures applied to the scores obtained from the test forms measuring the same construct, it is possible to make these scores interchangeable regardless of when and to whom these test forms are applied (Kolen & Brennan, 2004; Dorans & Holland, 2000). Test equating is a statistical and psychometric technique used for the adjustment of scores from different tests measuring the same construct in order to compare scores obtained from various forms of that test (Dorans & Holland, 2000; Skaggs, 2005). Felan (2002) points out that the scores obtained from different tests can be placed on a single scale and compared simultaneously via the statistical relationship established between the scores obtained from two

*Corresponding Author E-mail: meltem.yurtcu@gmail.com cguzeller@gmail.com

ISSN: 2148-7456 online /© 2018 DOI: 10.21449/ijate.316420

different forms measuring the same construct. According to a definition by Angoff (1971), test equating is the process of converting the unit scale of a test form to the unit scale of another test form. Kim and Hanson (2002) express equating as interchangeability of test forms after procedures applied to points from these test forms. In principle, the process of establishing the relationship between raw or scaled points used in two or more test forms is described as equating (Skaggs & Lissitz, 1986). The conditions required to be able to do equating are measuring the same construct, having equal reliability, equity, and invariance between groups (Dorans & Holland, 2000; Lord, 1980; Swaminathan & Gifford, 1983).

The right decision making end of these large scale exams that are extremely important for societies depends on reliability and validity of exams. Especially in equating of large-scale, there are a lot of situation that threaten reliability and validity. The some of the situations stem from multiple sources including measurement error, sampling error, measurement disturbances and administrative challenges. Measurement error usually refers to inaccurate associated with a measuring instrument (Wu, 2010). Depending on the equating method and pattern, the error emerging as a result of equating is of two types: random and systematic (Kolen, 1988; Felan, 2002). While random error that stems from answerer sampling is defined as standard error of equating (Kolen & Brennan, 2004); the other type of equating error, which is also known as equating bias, stems from violation of axioms or from biasedness (Zeng, 1991). Biasedness arises as a result of evaluation of an item with differential item functioning (DIF) by specialist opinion and involves sensitivity and differential item functioning analysis (Hambleton, 2006; Sireci & Mullane, 1994; ETS, 2009).

DIF surfaces as individuals with similar ability level but are in different subgroups differ in their probability for answering test items (Osterlind, 1983; Zumbo, 1999). Differential item function is of two types: uniform and non-uniform. It is considered uniform if the probability an item being answered correctly contains DIF in favor of a specific group for all ability levels but non-uniform if it contains DIF in favor of different groups at different ability levels (Zumbo, 1999). Investigation of differential item functioning (DIF) is with outmost important on the accuracy of the decisions taken as a result of large-scale examinations for societies when comparing measures across different groups (Lai, Teresi & Gershon, 2005; Swaminathan & Rogers, 1990). The presence of a DIF item(s) in the test, an indication of bias, will cause the obtained scores to be misleading (Zieky, 2002; Osterlind, 1983).

In the context of this study, the aim is to investigate the effect on the equating error obtained from the IRT-based equating methods according to the test containing DIF items and the number of DIF items in two tests with the same item parameters during the process of placing the points obtained from these tests on the same scale. Equalization of tests containing DIF items with item response models takes place in the literature using different methods and conditions (Demirus, 2015; Huggins, 2014). However, differentiation of the number of DIF items and the distribution of DIF items in test forms which be equalized in common item design equal groups makes this work unique from other studies. In this respect it will be contribute to literature. In this line, the basic research question may be formulated as:

"What are the effects of the number of DIF items in tests and of the tests containing DIF items on the equating error during the process of placing two math tests measuring the same construct on the same scale?"

2. METHOD

In this study purposes to indicate the effect of the number of DIF items and the distribution of DIF items in these forms, which be equalized on equating error. This is a basic research study in essence since it investigates the effect of the number of DIF items present in forms on

equating error with respect to the forms including DIF items by using IRT equating methods on common item pattern in equal groups.

2.1. Data Collection

Here, the study was conducted on the data set generated from the 2013-2014 TEOG exam on the basis of the assumption that the tests were taken by individuals with equal ability. Two different math test forms were generated with Wingen2 program by using item parameters in the math test of this exam. These forms are comprised of a medium-length test containing 15 common items aside from a set of 40 parallel questions. Hence, scores obtained from two tests containing 55 items per each were on the same scale. The item parameters of the math test were 0.20-0.76 for parameter a, 0.34-0.83 for parameter b, and 0.25-0.40 for parameter c. The common item pattern in equal groups was used as a pattern in equating. The forms A and B with 40 items per each were generated for different conditions in accordance with the three-parameter logistic model scored as 1-0 regarding the Item Response Theory models. Since the common form was so as to reflect A and B tests, it was generated by using the same parameters. The forms were generated to measure the same construct unidimensionally. For the ability distribution of the groups taking these forms, 1000 answers with normal distribution were generated so as the mean is 0 and standard deviation is 1. There are items with uniform DIF at B (medium) level in the common test and in the basic test on the generated forms. The DIF items were obtained as in favor of single group (in favor of males in TEOG); sizes of focus and reference groups are equal.

In order to answer the research question, six different conditions were considered: two different situations for number of DIF items (5 and 10) and three different situations for the test form containing the DIF items (form A, form B, and the Common form). The patterns of conditions are given in Table 1.

Table 1. The conditions determined with respect to the number of DIF items on forms and on the forms containing DIF items.

	Number of Items	**Total of 5 DIF Items**			**Total of 10 DIF Items**		
Form A	40	5 DIF Items	3 DIF Items	-	10 DIF Items	5 DIF Items	-
Form B	40	-	-	-	-	-	-
Common Form	15	-	2 DIF Items	5 DIF Items	-	5 DIF Items	10 DIF Items
		Condition 1	**Condition 2**	**Condition 3**	**Condition 4**	**Condition 5**	**Condition 6**

As it is seen in Table 1, six different conditions were obtained on the basis of different number of DIF items contained and the test forms these DIF items were on after forms A and B were generated as basis. Attention was paid to not to place the DIF items on tests consecutively.

2.2. Data Analysis

The common form was included in scores as internal anchor test in the study. Since the data belonging to test forms used in this study display similar difficulty and selectivity means, horizontal equating was done among these test form. The same parameters were used for common form data.

Separate conjecture methods were used for equating pattern used. PARSCALE 4.1 program was used for conjecture of parameters, IRTEQ program was used for test equating and

scaling. Data derivation and equating process were repeated 25 times for each condition and each method.

The root mean square deviation (RMSD) value was used in equating the test scores that the individuals with same ability level have received from different test forms. The RMSD values obtained from Mean-Mean, Mean-sigma, Stocking-Lord, Heabara equating methods were obtained by averaging 25 repeats.

3. FINDINGS

The six conditions were considered for the comparison of the equating error obtained by different IRT equating methods on the basis of the number and distribution of DIF items. In order to compare the condition as criteria, the equating errors in condition where both test forms do not contain DIF items.

Firstly, the condition where the 7^{th}, 12^{th}, 23^{rd}, 26^{th}, and 37^{th} items in the first 40 questions of the basic test, which is called test A and is among the math test to be equated, display uniform DIF with a difference of 0.6 at B level and there is no DIF item in the first 40 questions of the common test and form B was considered. This condition where there are five DIF items in the basic test and no DIF items in common test and form B is called Condition 1.

Condition 2 was created where DIF items are present both in the common test and the basic test, as number of DIF items is kept same. Under this condition, it is assumed that there are three DIF items, the 5^{th}, 17^{th}, and 33^{rd} items, in the first 40 questions of the basic test; and there is DIF in the 47^{th} and 53^{rd} items of the common test.

Condition 3 was created to analyze the RMSD value where DIF items are present only in the common test, as number of DIF items is fixed. Under this condition, it is assumed that there is DIF in the 51^{st}, 52^{nd}, 53^{rd}, 54^{th}, and 55^{th} items only in the common test form of the math test.

In order to investigate the effect of the change in the number of DIF items on equating error, the number of DIF items in the first 40 questions of the basic test is considered to be ten. Items that were considered as having DIF are the 5^{th}, 7^{th}, 12^{th}, 17^{th}, 23^{rd}, 26^{th}, 29^{th}, 33^{rd}, 37^{th}, and 40^{th} items. The condition where there is no DIF item in the first 40 questions of the common test and form B is called Condition 4.

Condition 5 was created which tests the DIF items are present in while the number of DIF items in tests to be equated is taken as ten and the number of DIF items is fixed. For this condition, it is assumed that the 7^{th}, 12^{th}, 23^{rd}, 26^{th}, and 37^{th} items of the first 40 questions on A test and the 51^{st}, 52^{nd}, 53^{rd}, 54^{th}, and 55^{th} items of the common test have DIF.

Created condition 6 where there are ten DIF items only in the common test is assumed that only the 46^{th}, 47^{th}, 48^{th}, 49^{th}, 50^{th}, 51^{st}, 52^{nd}, 53^{rd}, 54^{th}, and 55^{th} items on the common test form have DIF.

We examined RMSD equating errors of equating done by four methods for 6 conditions and math test forms without DIF as scaling method. The equating errors, which were obtained as the points taken from tests A and B belonging to these conditions were placed on same scale, were investigated with respect to IRT equating methods. These values were shown in Table 2.

Table 2. The RMSD equating errors of equating done by four methods for conditions where math test forms without DIF.

	Mean- Mean	Mean-Sigma	Haebara (HB)	Stocking-Lord (S-L)
The equating errors for test forms without DIF	0.057616	0.179619	0.17014	0.171374
Condition 1	1.14101	0.842776	0.98555	0.597466
Condition 2	0.348804	0.511489	0.328713	0.295562
Condition 3	0.39065	0.588079	0.308391	0.291414
Condition 4	1.165186	0.886565	0.600028	0.606109
Condition 5	0.646586	0.915705	0.546247	0.519187
Condition 6	0.318883	0.69995	0.352803	0.332708

condition 1: five DIF items in the test A and no DIF items in common test and form B
condition 2: five DIF items in the test A and two DIF items in test B
condition 3: five DIF items in the common test of the math forms
condition 4: ten DIF items in test A and there is no DIF in the common test and form B
condition 5: ten DIF items in test A and five DIF items test B
Condition 6: ten DIF items in the common test of the two math forms

When the tests forms don't include DIF items, the lowest error among the IRT equating methods looks to be with Mean-Mean method. It is followed by the equating error calculated by the Haebara method. The highest error was obtained by Mean-sigma method.

In condition 1, the lowest error among the IRT equating methods looks to be with Stocking-Lord method in conditions B. It is followed by the equating error calculated by the Mean-sigma method. The highest error was obtained by Mean-Mean method.

In condition 2, condition 3 and condition 5 the lowest error among the IRT equating methods looks to be given by Stocking-Lord method. It is followed by the equating error calculated by the Heabara method, one of the characteristic curve methods. The highest error was produced by Mean-sigma method in this condition.

When condition 4 is examined, the lowest error among the IRT equating methods looks to be given by Haebara method. Following this method, the points obtained by the Stocking-Lord method look to have the next lowest error. It is observed that the highest error was obtained by Mean-sigma method.

When Condition 6 is examined, the lowest error among the IRT equating methods looks to be given by Mean-Mean method. The error coefficient obtained by the Heabara method follows. It is observed that the highest error was obtained by Mean-sigma method.

4. RESULTS AND CONCLUSION

Changes in the curriculum, such as test structure, test length, and retention exposure can create bias among individuals (Stocking & Lewis, 1998). The presence of questions, which may create a bias in favor of a specific group in one or two of the tests being equated, will affect the validity of this test (Osterlind, 1983; Zieky, 2002). It is also important to test whether the anchors items included in the test have DIF (Klein & Jarjoura, 1985; Cook & Petersen, 1987).

In accordance with the purpose of the study, it was investigated that inclusion of the DIF items in test equating process casts doubt on the accuracy of the scores generated as a result of equating. RMSD was used as the criteria value because of providing an estimate by combining the random and systematic equating error (Puhan, 2010; Sinharay & Holland, 2007) and these RMSD values of IRT equating methods were considered were compared to each other. Variations in the RMSD value, which was considered as the equating error, were examined with

respect to the number of DIF items and with respect to which test forms have the DIF items among the tests to be equated.

Presence of DIF items in any of tests to be equated causes a decrease in errors calculated by all IRT equation methods. While increasing the number of DIF items only in test A causes an increase in errors for all methods except for Haebara method, increasing the number of the DIF item only in common test causes increase in errors for all methods except for the mean- mean method. Increase in the number of DIF items both in the common test and the basic test causes an increase in error calculated by all methods. When conditions that include the same number of DIF items in common test are compared, the presence of DIF items in the basic test also increases the error.

That there are DIF items in both tests causes it to have less error than the condition where only test A has DIF items except for mean-sigma method in competing condition 5 and condition 4. To see this, it can be compared to condition 1 and condition 2; condition 1 and condition 3; condition 4 and condition 6.

When it is examined all conditions including condition where both test forms do not contain DIF items, generally it can be seen that lowest equation errors are obtained by Stocking-Lord method and the highest error was obtained by Mean-sigma method during equating done in the study.

According to research studies that have a common finding is that item characteristic curve methods give more accurate than moment methods (Beguin, 2002; Kim & Cohen, 1992; Way & Tang, 1991; Stocking & Lord, 1983; Ogasawara, 2001). Kilmen and Demirtaşlı (2012) also express their study that equation errors are obtained by Stocking-Lord method indicate less errors than other IRT methods. The c parameter is never considered in the calculation of the scale factor since the mean-sigma and mean-mean methods derive the scaling factors from the descriptive statistics of the distribution of b-parameters. We may state that the equating error obtained by Mean-Mean and mean-sigma methods is higher due to added DIF items being uniform and being a result of a change of 0.6 unit at B level.

In the literature, there is very little work that compares the methods of equalization on this subject. Demirus (2015), who examines the effects of items with DIF on the real data, in case the anchor items display uniform DIF for a group, the mean-mean method produces the largest error, the mean-sigma method yields the smallest. On the anchor items without DIF the biggest equating error has been obtained by mean-sigma method and smallest equating error has been obtained by Stocking-Lord and Haebara methods. This is partly similar to our findings.

In future studies, the status of mixed-structure test that includes DIF items can be examined. The DIF level taken the uniformly in this study can be considered at many different levels. In addition, as a different dimension of this study, it is possible to examine how the results will be observed when the skill levels of the groups receiving the tests to be equal are different.

5. REFERENCES

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

Béguin, A. A., Hanson, B. A. & Glas, C. A. W. (2000). Effect of unidimensionality on separate and concurrent estimation in IRT equating. Paper presented at the *Annual Meeting of the National Council on Measurement in Education*, New Orleans, LA. Available from http://www.bah.com/papers/paper0002.html

Cook, L. L. & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11*, 225–244

Demirus, K. B. (2015). *Ortak maddelerin değişen madde fonksiyonu gösterip göstermemesi durumunda test eşitlemeye etkisinin farklı yöntemlerle incelenmesi.* Doktora Tezi, Ankara: Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü.

Dorans, N. J. & Holland, P. W. (2000). Population invariance and the equatability of tests: basic theory and the linear case. *Journal of Educational Measurement*, *37* (4), 281- 306.

Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41*, 43-68.

Educational Testing Service. *Guidelines for fairness review of assessment*. Retrieved May 22, 2015 from http://www.ets.org/Media/About_ETS/pdf/overview.pdf

Felan, G. D. (2002). Test Equating: Mean, Linear, Equipercentile and Item Response Theory. Paper presented at the *Annual Meeting of the South West Educational Research Association*, Austin.

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care. 44*(11), 182-188.

Huggins, A. C. (2014). The effect of differential item functioning in anchor items on population invariance of equating. *Educational and Psychological Measurement. 74*(4), 627-658.

Kilmen, S. & Demirtaslı, N (2012). Comparison of test equating methods based on item response theory according to the sample size and ability distribution. *Procedia - Social and Behavioral Sciences, 46*, 130-134.

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

Kim, S. & Cohen, A.S. (1992). Effects of linking methods on detection of DIF. *Applied Psychological Measurement*, *29*(1), 51-56.

Kim, S., Walker, M.E. & McHale, F. (2010). Comparisons among designs for equating mixed-format tests in large-scale assessments. *Journal of Educational Measurement, 47* (1), 36-53.

Klein, L. W. & Jarjoura, D. (1985). The importance of content representation for common item equating with non-random groups. *Journal of Educational Measurement, 22*, 197-206.

Kolen, M. J. (1988). Traditional equating methodology. *Educational Measurement Issues and Practice, 7* (4), 29-36.

Kolen, M. J. & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking* (2nd edition). USA: Springer.

Lai, J. S., Teresi, J. & Gerson, R. (2005). Procedures for the analysis of differential item functioning (DIF) for small sample sizes, *Evaluation & The Health Professions, 28*(3), 283-294.

Lord, M. F. (1980). *Application of Item Response Theory to Practical Testing Problems.* New Jersey: Lawrence Erlbaum Associates Publishers.

Ogasawara, H. (2001). Item response theory true score equating and their standard errors. *Journal of Educational Behavioral Statistics*, *26*(1), 31-50.

Osterlind, J. S. (1983). *Test item bias*. London Sage Publications.

Puhan, G. (2010). A comparison of chained linear and post stratification linear equating under different testing conditions. *Journal of Educational Measurement, 47*(1), 54–75.

Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures, *Journal of Educational Measurement, 27*(4). 361-370.

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.

Sireci, S.G. & Mullane, L. A. (1994). Evaluating test fairness in licensure testing: The sensitivity review process. *CLEAR Exam Review*. *5*(2), 22-27.

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

Stocking, M.L. (1988). *Factors affecting the sample invariant properties of linear and curvilinear observed- and true- score equating procedures*. (ETS Research Report NO. RR-88-41). Princeton, NJ: Educational Testing Service.

Stocking, M.L. & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 23*, 57-75.

Swaminathan, H. & Gifford, J.A. (1983). Estimation of parameters in the three parameter latent trait model. In D. Weiss (Ed.), *New horizons in testing.* New York: Academic Press.

Zeng, L. (1991). Standard errors of linear equating for the single-group design (ACT Research Report 91-4). Iowa City, IA: American College Testing.

Zieky, M. (2002). Ensuring the fairness of Licensing Tests. *CLEAR Exam Review*. *12*(1), 20-26.

Zumbo, B.D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores*. Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Way, W. D. & Tang, K.L. (1991). A comparison of four logistic model equating methods. Paper presented at the *Annual Meeting of the American Educational Research Association*, Chicago.

Wu, M. (2010). Measurement, Sampling, and Equating Errors in Large-Scale Assessments. *Educational Measurement: Issues and Practice, 29* (4), 15–27.