Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 46 (2012) 130 – 134

Procedia
Social and Behavioral Sciences

WCES 2012

Comparison of test equating methods based on item response theory according to the sample size and ability distribution

Sevilay Kilmen*, Nukhet Demirtasli**

*Abant Izzet Baysal University, Bolu, Turkey
* Ankara University, Ankara, Turkey

Abstract

In this study, it was aimed that equating methods of "mean-mean", "mean-sigma", "Heabera" and "Stocking-Lord" were compared in terms of the ability distribution and sample size variables based on Item Response Theory (IRT). Common item nonequivalent groups equating design was used. In this study 600 dichotomously responded data matrices were generated. Ability parameters of have been estimated from simulated data with expected a posteriori method. Results showed that, in the case of groups having similar and different ability distribution with different sample size (500 and 1000), test equating application with "Stocking-Lord" method gave less equation error has been obtained.

Keywords: Item Response Theory; test equating; common item nonequivalent groups equating design.

1. Introduction

Test equating is a statistical process providing the interconversion of the scores obtained from different test forms measuring the same structure. Equalized scores have the same meanings regardless of when and to whom the test was applied. Therefore, scores obtained from a test form can be compared to the scores obtained from another test form (Kolen and Brennan, 2004). In the event that the students are simultaneously given different forms of a test and the difficulty level between the forms are not equal, it is possible that the individuals given the difficult test get lower scores than the individuals given the easy test. This situation makes it difficult to compare the points obtained from different test forms. Test equating prevents the possible unfairness against individuals given the difficult test and eliminates the bias problems caused by the difference in the difficulty levels of the test forms (Angoff, 1971; Cook and Eignor, 1991; Hambleton, Swaminathan and Rogers, 1991).

Test equating design frequently used in test equating studies is "common items non-equivalent equating design". Items in the both forms (such as A and B form) of a test are referred as anchor/common items. In this design, there are two different group tests (A, B). Each group responds to one of the forms. In this design, anchor items are used in the placing of the item and the ability parameters on the same scale. An equation that will provide the interconversion of the scores to be obtained from two test forms is created. B equation has an equation constant and a curve (Cook and Eignor, 1991; Vale, 1986).

1.1. Test Equating Methods based on Item Response Model

1.1.1. ***Mean-mean method.*** It is a method developed by Loyd and Hoover (1980) and it is used to convert the score scale of the difficulty and discrimination parameters. In this method, average of the discrimination parameter is used in the determination of the curve; and average of the difficulty parameter is used in the determination of the equation constant.

 doi:10.1016/j.sbspro.2012.05.081

1.1.2. ***Mean-sigma method.*** It is a score converting process defined by Marco (1977). In this method, standard deviation is used in the determination of the equating equation curve; and average difficulties of the tests are used in the determination of equation constant as in the case of mean-mean method.

1.1.3. ***Haebara method.*** In this method, equation curve and equation constant are obtained by using the difference between the item characteristic curves. This approach is developed by Haebara (1980) who named after the method. For the participants having a certain ability level, the difference between the item characteristic curves is the sum of the squares of the item characteristic curves belonging to each item. Equation constant and equation curve minimizing this difference are tried to be found (Kolen and Brennan, 2004; Raju and Arenson, 2002).

1.1.4. ***Stocking-Lord method.*** In this method, equation curve and equation constant is obtained by using the difference between the item characteristic curves. Unlike Haebara approach, in Stocking-Lord (1983), for the participants having a certain ability level, the difference between the item characteristic curves is the square of the sum of the difference between the item characteristic curves belonging to each item. Equation constant and equation curve minimizing this difference are tried to be found (Kolen and Brennan, 2004).

In this study, test equating methods were compared according to the similar and different ability

distribution and sample size (500, 1000) variables based on Item Response Theory (IRT). It is believed that the results of this research will be helpful to the researchers, teachers and measurement and evaluation associations/centers carrying out large scale testing implementations for the selection of the suitable equating method in the equating of different test forms. It is also believed that this study will also contribute to the theoretical researches carried out in order to determine the conditions necessary to obtain equating results having the least errors.

2. Purpose of Study

In this study, it was aimed that equating methods of "mean-mean", "mean-sigma", "Heabera" and "Stocking-Lord" were compared according to the ability distribution (similar and different ability distribution) and sample size (500, 1000) variables based on Item Response Theory. Common item nonequivalent groups equating design on dichotomous simulated data adapted to 3-parameters IRT model was used.

3. Methods

In this research, WinGen2 simulation data generating program has been used in creating the research data. In the first step of data generating, 2 simulative test forms which are called as K and L, have one-dimensional feature and are in conformity with 3-parameter logistic models have been created. These forms are the test forms to be equated. K and L forms are consisted of 50 (40 + 10) items per each. This research has been planned based on the common item non-equivalent group equating design. For that reason, first 40 items included in the K and L forms are different from each other. Other 10 items are the same in other two test forms. Data in these test forms are organized in a way that they have item distribution having 0 average in terms of item difficulties (b). Discrimination parameter (a) varies between 1-2 values.

In the second step of data generating, 6 groups have been created in terms of sample size and ability distribution and those groups are;

- *Two groups* having 0 average for 500 people and 1 standard deviation [N(0,1)]
- *One group* having 1 average for 500 people and 1 standard deviation [N(1,1)]
- *Two groups* having 0 average for 1000 people and 1 standard deviation [N(0,1)]
- *One group* having 1 average for 1000 people and 1 standard deviation [N(1,1)].

Four different equating conditions with 6 groups of data have been created in terms of sample size and ability distribution and those groups are;

- Test equating in 500-people groups having similar ability levels

- Test equating in 500-people groups having different ability levels
- Test equating in 1000-people groups having similar ability levels
- Test equating in 1000-people groups having different ability levels

100 replications have been made for each of the 6 data groups and in total they are equal to 600 replications. Therefore, in total 600 data of 1-0 have been obtained in order to be used in the research. BILOG-MG 3.0 (Zimowski, Muraki, Mislevy and Bock, 1996) program has been used in the estimation of the item and ability parameters belonging to these data. With the help of IRTEQ (Han, 2007) test equating program, equation equity between the two different forms (K and L) of a test has been created. In the comparison of the equating errors obtained from test equating methods " root mean square deviation" (RMSD) has been used as a criterion. When this value is low it means that equations with less error are made.

4. Findings and Results

Findings regarding the error amounts obtained from mean-mean, mean-sigma, Haebara and Stocking-Lord test equating methods under the conditions of different sample size (500-1000) and ability distribution ([N(0,1)-N(0,1)] and [N(0,1)-N(1,1)]) have been shown in the Table-1:

Table 1. RMSD Values Obtained from Test Equating Methods under the Conditions of Different Sample Size and Ability Distribution

N	Ability distribution	Mean-mean	Mean-sigma	Haebara	Stocking-Lord
500	[N(0,1)-N(0,1)]	0,120	0,111	0,096	**0,091**
500	[N(0,1)-N(1,1)]	0,170	0,240	0,168	**0,150**
1000	[N(0,1)-N(0,1)]	0,090	0,098	0,086	**0,077**
1000	[N(0,1)-N(1,1)]	0,126	0,189	0,120	**0,117**

When the Table 1 is analyzed, it has been observed that the test equating method giving the less error in the equating of two test forms which are in conformity with 3-parameter logistic model is "Stocking-Lord" method for the groups having 500-people similar ability distribution and 1000-people similar and different ability distributions. This method respectively follows Heabara and mean-sigma methods. For the aforesaid condition, the method having the most errors is "mean-mean" method. For the aforesaid conditions, the method having the most errors is "mean-sigma" method.

It has been determined that in groups having 500-people different ability distribution, the most suitable test equating method to be used in the equating of the different test forms which are in conformity with 3-parameter logistic model is "Stocking-Lord" method developed based on the characteristic curve approach. This method respectively follows Heabara and mean-mean methods. For the aforesaid condition, the method having the most errors is "mean-sigma" method.

Findings regarding the comparison of the equating errors estimated from test equating methods in terms of sample size and ability distribution have been shown in the Figure 1.

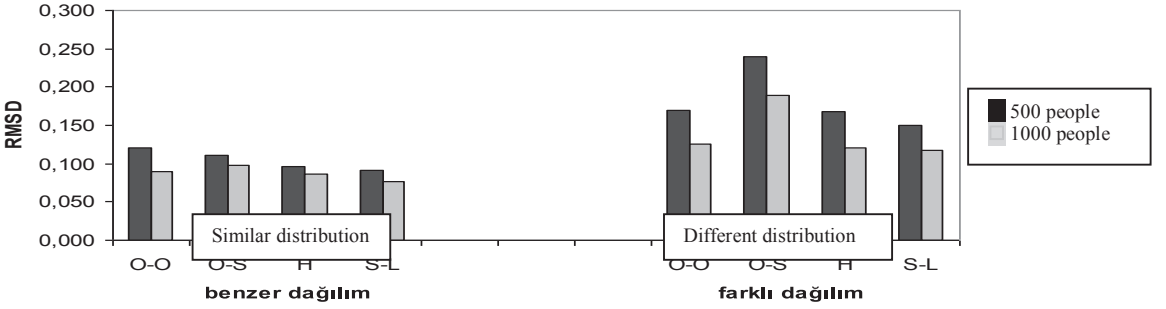

Figure 1. Errors estimated from Test Equating Methods in terms of Sample Size and Ability Distribution

When Figure 1 is analyzed, it has been observed that for both similar and different distributions, RMSD averages obtained from 500-people groups have higher values than the 1000-people samples. In other words, it has been observed that as sample size increases, RMSD averages decreases. When Figure 1 is analyzed in terms of ability distribution, it has been found out that for both 500-people group and 1000-people group, the groups having similar ability distribution have less equating errors than different ability distributions in terms of ability distribution variable.

Based on the findings of this research, it can be stated that within the conditions used in this research, the best equating condition can be obtained by using Stocking-Lord equating method in the groups having 1000-people sample size and similar ability distribution.

5. Conclusions

For the groups having both similar ability distribution and different ability distribution it has been determined that at 500 and 1000-people sample sizes there are less equating errors than the other equating methods of Stocking-Lord method. Mean-mean method and mean-sigma method have more equating errors than the characteristic curve methods in all conditions. We can encounter this finding in the researches carried out by Stocking and Lord (1983), Baker and Al-Karni (1991), Cohen and Kim (1998), Hanson and Beguin (2002), Kim and Kolen (2006), Hung, Wu and Chen (1991), Way and Tang (1991), Karkee and Wright (2004), Kaskowitz and De Ayala (2001), Kim and Lee (2004), Kim and Lee (2006), Kim and Kolen (2004), Kim and Song (2004) and Ogasawara (2001).

In this research when the findings are analyzed in terms of sample size it has been determined that RMSD value average obtained from 1000-people groups are less than the RMSD values obtained from the groups having 500-people sample size. As the sample increases, equating error decreases. In the researches carried out by Yotinprasert (1986), Speron (2009), Harris (1993), Hanson and Beguin (2002), Kim and Cohen (2002) and Yang (1997), Bastari (2000), Tate (2000), Çetin (2009) and Lee & Ban (2010) similar result has been obtained. Unlike these findings, in a research where equating based on 1-parameter model has been made and which has been carried out by Suanthong (1998), it has been observed than between the 100, 300, 500-people sample sizes used in the research, minimum equating errors have been seen in 100-people sample sizes.

When test equating methods have been compared according to ability distribution variable, it has been determined that RMSD value average obtained from equating studies belonging to the groups having similar ability distribution are lower than the RMSD values obtained from the groups having different ability distributions. In other words, as the ability distributions of the groups to be equated become similar to each other, RMSD value averages decrease. In the researches carried out by Hanson and Beguin (2002), Kim and Cohen (2002), Bastari (2000) and Kim and Lee (2006) similar result has been obtained.

6. Recommendations

Based on the findings of this research, it can be suggested that Stocking-Lord test equating method is used for the test equating studies in practical and 1000-people sample size is preferred instead of 500.

In this research, simulative data graded as 1-0 based on Item Response Model has been used for sample groups of 500 and 1000-people. These equating methods can be also compared on the actual data where the grading in 1-0 and multiple categories are carried out together or which is graded as multiple categories at different sample sizes. This research has been limited with mean-mean, mean-sigma, Haebara and Stocking-Lord methods. Errors belonging to the concurrent equating methods of the parameters can be researched at different conditions. In this research, sample size and ability distribution which are thought to affect the test equating errors have been used. As well as these variables, effects of the different variables such as anchor item length, test length, computer program used in ability estimation etc. on equating errors can also be studied.

References

Angoff, W. H. (1971*). "Scales, Norms and Equivalent Scores".* In Thorndike, R. L. (Ed.), Educational Measurement (p. 509-600). Washington: American Council on Education.

Baker, F.B. and Al-Karni, A. (1991). A Comparison of Two Procedures for Computing IRT Equating Coefficients. *Journal of Educational Measurement, 28* (2), 147-162.

Bastari, B. (2000). *Linking Multiple Choice and Constructed Response Items to a Common Proficiency Scale.* Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Cohen, A. S. and Kim, S. H. (1998). An Investigation of Linking Methods Under the Graded Response Model. *Applied Psychological Measurement, 22* (2), 116-130.

Cook L. L. and Eignor R. E. (1991). An NCME Instructional Module on IRT Equating Methods. Instructional Topics in Educational Measurement. Educational Measurement: *Issues and Practice, 10* (1), 37–45.

Crocker, L. and Algina, J. (1986*). Introduction to Classical and Modern Test Theory.* New York: Holt, Rinehart and Winston.

Çetin, E. (2009). *Dikey Ölçeklemede Klasik Test ve Madde Tepki Kuramına Dayalı Yöntemlerin Karşılaştırılması.* Yayımlanmamış yüksek lisans tezi, Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Embretson, S. E. and Reise, S. P. (2000*). Item Response Theory for Psychologists.* London: Lawrence Elbaum Associates, Publishers.

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. and Rogers, H. (1991). *Fundamentals of Item Response Theory*. Newbury Park CA: Sage.

Han, K. T. (2007). IRTEQ: Windows Application That Implements IRT Scaling and Equating [Computer Program]. Web: http://www.umass.edu/remp/software/irteq adresinden 5 May 2008'de alınmıştır.

Hanson, B. A. and Béguin, A. A. (2002). Obtaining a Common Scale for Item Response Theory Item Parameters Using Separate Versus Concurrent Estimation in the Common-Item Equating Design. *Applied Psychological Measurement, 26* (1), 3–24.

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

Hung, P., Wu, Y., and Chen, Y. (1991*). IRT Item Parameter Linking: Relevant Issues for the Purpose of Item Banking.* Paper presented at the International Academic Symposium on Psychological Measurement, Taiwan.

Karkee, T. B. And Wright, K. R. (2004, April*). Evaluation of Linking Methods for Placing Three Parameter Logistic Item Parameter Estimates Onto Rasch Scale.* Paper presented at the Meeting of the American Educational Research, San Diego, California.

Kaskowitz, G. S. and De Ayala, R. J. (2001). The Effect of Error in Item Parameter Estimates on the Test Response Function Method of Linking. *Applied Psychological Measurement, 25* (1), 39-52.

Kim, S. H. and Cohen, A. S. (2002). A Comparison of Linking and Concurrent Calibration Under the Graded Response Model. *Applied Psychological Measurement, 26* (1), 25–41.

Kim, S. and Kolen, M. J. (2006). Robustness of Format Effects of IRT Linking Methods for Mixed Format Tests. *Applied Measurement in Education, 19* (4), 357-381.

Kim, S., and Lee, W. (2004). IRT Scale Linking Methods for Mixed-Format Tests (ACT Research Report 2004–5). Iowa City, IA: ACT, Inc.

Kim, S. and Lee, W. C. (2006). An Extension of Four IRT Linking Methods for Mixed-Format Tests. *Journal of Educational Measurement, 43* (1), 53–76.

Kim, S., and Song, M.-Y. (2004). *Least Squares Estimation of IRT Scale Linking Coefficients under the Graded Response Model*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Kolen, M. J. and Brennan, R. L. (2004*). Test Equating, Scalling, and Linking*. (second edition). USA: Springer.

Lee, W. C. and Ban, J. C. (2010). A Comparison of IRT Linking Procedures. *Applied Measurement in Education, 23* (1), 23-48.

Loyd, B. H. and Hoover, H. D. (1980). Vertical Equating Using the Rasch Model. *Journal of Educational Measurement, 17 (*3), 179-193.

Marco, G. L. (1977). Item Characteristic Curve Solutions to Three Intractable Testing Problems. *Journal of Educational Measurement, 14* (2), 139-160.

Ogasawara, H. (2001). Standard Errors of Item Response Theory Equating/Linking by Response Function Methods. *Applied Psychological Measurement, 25* (1), 53–67.

Raju, N. S., and Arenson, E. A. (April, 2002). *Developing a Common Metric in Item Response Theory: An Area-Minimization Approach.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Stocking, M. L. and Lord, F. M. (1983). Developing a Common Metric in Item Response Theory. *Applied Psychological Measurement, 7* (2), 201-210.

Speron, E. (2009). *A Comparison of Metric Linking Procedures in Item Response Theory.* Unpublished doctoral dissertation, Illinois Institute of Technology, Chicago, Illinois.

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.

Vale, C. D. (1986). Linking Item Parameters Onto a Common Scale. *Applied Psychological Measurement, 10* (4), 333-344.

Way, W. D., and Tang, K. L. (1991, April 4-6). *A Comparison of Four Logistic Model Equating Methods.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Yang, W. L. (1997*). The Effects of Content Homogeneity and Equating Method on the Accuracy of Common Item Test Equating.* Unpublished doctoral dissertation, Michigan State University, Michigan.

Yotinprasert, S. (1986*). The Effect of Sample Size on Error Produced by Tucker and Rasch Equating Methods under Common Items Nonrandom Grups Design.* Unpublished doctoral dissertation, Florida State University, Tallahassee.

Zimowski, M. F., Muraki, E., Mislevy, R. J. and Bock, R. D. (1996). *BILOG– MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items.* Chicago: Scientific Software International.