Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 174 (2015) 2075 – 2083

Procedia
Social and Behavioral Sciences

A comparison of IRT-based methods for examining differential item functioning in TIMSS 2011 mathematics subtest

Burhanettin Özdemir*

Hacettepe University, Department of Educational Measurement and Statistics, Ankara,06800,Turkey

Abstract

The purpose of this study is to determine items which have differential item functioning (DIF) in TIMSS 2011 mathematics subtest with three different item response theory (IRT)-based DIF methods and compare results of these methods. For this purpose, DIF values obtained by Lord's Chi - Square , Raju's Area and Likelihood-Ratio Test methods were compared with respect to gender (males were the reference group while females were the focal group) to test whether these procedures yielded similar results. In addition, item purification was performed for each methods and results were compared in order to determine the effect of item purification. These comparisons can provide evidence for determining the best models for detecting DIF items. Results indicated that 2PL IRT model fitted best to the data for both Lord's Chi - Square method and Raju's Signed Area method. Although number of items detected as DIF differed for each methods, 2 out of 22 dichotomous items in the test observed consistently across all methods, which were more likely to be answered correctly by males after controlling for overall ability.

Peer-review under responsibility of the Sakarya University

Keywords: Item Response Theory, Differential Item Functioning, TIMSS

1. Introduction

International assessments such as the Third International Mathematics and Science Study (TIMSS) reveal students' achievement level in science and mathematics and also to get information about the effectiveness of the present school curricula in participating countries (Keser, 2005; Uzun, Butuner & Yigit, 2010). However, some of the result of TMSS could be biased with respect to students' achievement and effectiveness of the education system

* Corresponding author. Tel.: +90-542-646-49-36; fax: +90-312-297-85-66.
E-mail address: b.ozdemir@hacettepe.edu.tr

Peer-review under responsibility of the Sakarya University
doi:10.1016/j.sbspro.2015.02.004

of participating countries. One of the main reason behind these unexpected results is that items in the test may function differently with respect to gender and cultural differences.

A widely accepted definition of differential item functioning (DIF) was that an item is identified as DIF if examinees of equal ability, but from different subgroups do not have an equal probability of correctly responding to that item (Hambleton & Rogers, 1989). If the discrepancy in item performance between the subgroups of interest is equal across the entire range of abilities then the DIF is said to be "uniform". However, if the difference between the subgroups is not consistent across the entire range of abilities then the DIF is said to be "non-uniform" (Hambleton, Clauser, Mazor & Jones, 1993)

DIF items can lead to biased measurement of ability because the measurement is affected by so-called *nuisance factors* (Ackerman, 1992). It is important to clarify one concept which has been used previously instead of DIF but now has another meaning; item bias (Scheuneman & Bleistein, 1997). A biased item displays DIF; however that is not sufficient for the item being biased. DIF is a statistical property of an item while item bias is more general and lies in the interpretation (Camilli & Shepard, 1994; Clauser & Mazor, 1998; Wilberg, 2007). An observed difference does not mean that there exists measurement bias since it might be a real difference in ability (Camilli, 2006). Item impact refers to when test takers from different groups have different probabilities of responding correctly to an item due to true differences in ability measured by the item (Dorans & Holland, 1993; Wilberg, 2007).

There are different methods which aim to determine DIF items and degree of DIF. The reasons for the differences in findings were posited to be due to the use of different criteria for identifying and flagging DIF, for example, measures of magnitude versus statistical significance (Borsboom, 2006; Hambleton, 2006; Millsap, 2006).

DIF methods are generally classified in to two groups, those methods based on item response theory (IRT) and those not based on IRT. For the IRT-based methods, the estimation of an IRT model is required, and a statistical testing procedure is followed, based on the asymptotic properties of statistics derived from the estimation results. For the latter, the detection of DIF items is usually based on statistical methods for categorical data, with the total test score as a matching criterion (Magis et al., 2010). In some research, IRT-based and Non-IRT-based methods are called as *parametric DIF methods* and *nonparametric,* respectively.

For dichotomously scored items, the usual IRT models are the logistic models with one, two, or three parameters. It was further denoted by 1PL, 2PL, and 3PL models, respectively. The 3PL model can be written as:

$$p\left(xi=1|\theta_n\right)=c_i+\left(1-c_i\right)\frac{\exp\left[a_i\left(\theta_n-b_i\right)\right]}{1+\exp\left[a_i\left(\theta_n-b_i\right)\right]} \tag{1}$$

where *xi* is the binary response of subject *n* to item *i*; θ_n is the ability of subject *n*; and a_i, b_i, and c_i are, respectively, the discrimination, difficulty, and pseudo-guessing parameters of item *i*. The 2PL model can be obtained from Equation 1 by fixing c_i to 0; the 1PL model can be obtained by additionally fixing a_i to 1. In this study, results of IRT-based DIF detection methods were compared and the 2PL IRT method was used to estimate ability and item parameters.

1.1 IRT-based DIF methods.

The basic idea of LRT is that item parameters should be invariant across different subgroups. In order to test item parameter invariance, likelihood of a compact model in which the parameters are constrained to be the same and an augmented model in which all variables of interest are allowed to vary between the subgroups are compared. The significance of this comparison is tested by means of the usual likelihood ratio test. Based on the selected IRT model, not only the item difficulties (1PL model), but also discriminations (2PL model), and pseudo-guessing parameters (3PL model) are allowed vary between the groups. The main idea is to compare the likelihood of two models and choose the model which has the largest likelihood. The LRT test statistic is defined as

$$G^2 = -2ln\frac{L\left(model\,a\right)}{L\left(model\,b\right)} = -2\left[\ell\left(c\right)-\ell\left(\mathrm{a}\right)\right] \sim \chi^2_{(m)} \tag{2}$$

where m is the difference in number of parameters between the augmented and the compact model.

The second IRT-based DIF method is called ***Lord's chi-square test*** (Lord, 1980) and is based upon the null hypothesis of equal item parameters in both subgroups and a statistic with a chi-square distribution under the null hypothesis (Magis,2010). Although three different item response models (1PL, 2PL, 3PL) can be fitted, before the analysis item parameters must be scaled with a common metric prior to statistical testing.. The *Qj* statistic is defined as;

$$\mathrm{Q}_j = \left(v_{jR}-v_{jF}\right)'\left(\sum jR-\sum jF\right)^{-1}\left(v_{jR}-v_{jF}\right) \tag{3}$$

where $V_{j\mathrm{R}}$ =($a_{j\mathrm{R}}$, $b_{j\mathrm{R}}$, $c_{j\mathrm{R}}$) and $V_{j\mathrm{F}}$ =($a_{j\mathrm{F}}$, $b_{j\mathrm{F}}$, $c_{j\mathrm{F}}$) are the vectors of item discrimination, difficulty, and pseudo-guessing estimates of item *j* in the reference group and focal group, respectively, and $\sum jR$ and $\sum jF$ are the corresponding variance–covariance matrices.

The third IRT-based DIF method is called the *Raju's Signded Area* method (Raju, 1988, 1990). In this method, the (signed) area between the item characteristic curves for the focal group and the reference group is computed and the corresponding *Z* statistic is based on the null hypothesis that the true area is zero. A common metric is required prior to the test. Any item response model can be considered with Raju's (1988) approach (Magis,2010).However, in this model the pseudo-guessing parameters for both groups of subjects are constrained to be equal. *Z* statistic for 1PL model is simply given as follows:

$$Z = \frac{b_{jR}-b_{jF}}{\sqrt{\hat{\sigma}^2_{jR}-\hat{\sigma}^2_{jF}}} \tag{4}$$

For 2PL and 3PL models, the formula for *Z* is much more complex and can be found in Raju (1990) (Magis et al, 2010).

1.2 Item purification

Item purification is based on iterative elimination of DIF items which prevent the inflation of Type-I error rate and increase the accuracy of the results. Especially, Type-I error inflates when DIF items are taken into account during the computation. As a result, more non-DIF items are incorrectly flagged as DIF (Clauser, Mazor, & Hambleton, 1993. Item purification iteratively removes the items currently flagged as DIF from the test scores to get purified sets of items, unaffected by DIF. With IRT based methods, item purification acts rather when item parameters in both groups of respondents are being rescaled, usually onto the reference group scale. At each step of the purification process, rescaling is made by removing all items currently flagged as DIF ((Magis and Fagon, 2013).

Item purification procedures are useful and powerful when there are only a few DIF items in the test. However, in the case of having to many DIF items in the tests, unwanted DIF items effect on non-DIF items may not be completely eliminated by item purification procedures.

2. Purpose

The purpose of this study is to determine items which functions differently with respect to gender of students in TIMSS 2011 mathematics subtest by three different item response theory (IRT)-based DIF methods and compare results of these methods. 2PL IRT method was used to estimate both ability and item parameters.

It is assumed that these three IRT-based techniques would show substantial agreement in the detection of DIF among the same set of mathematics subtest items, but vary in the number of items flagged with DIF due to different assumptions and criteria used.

3. Method

Real data from TIMSS 2011 mathematics subtest booklet 2, which was administered to 488 8th grade students (251 male and 237 female students), was used to evaluated three different DIF detecting methods. DIF values obtained by Lord's Chi - Square, Raju's Area and Likelihood-Ratio Test methods were compared with respect to gender, where males were the reference group and females were the focal group, to test whether these procedures yielded similar results. In addition, item purification was performed for each methods and results were compared in order to determine the effect of item purification on each methods. "difR" package in R software was used to conduct analysis for each methods.

4. Findings

Some required assumptions of methods have to be checked before conducting the analysis such as assumptions of IRT models. An underlying assumption of many IRT models is that the items within a scale are unidimensional, i.e., that a single underlying trait exclusively determines the probability of item responses (Embretson & Reise, 2000). While there are a number of different assumptions, methods, and software available to assess for dimensionality, such as assessing the fit of the data within Rasch models (Glas & Verhelst, 1995; Rasch, 1960; Rizopoulos, 2006,Yang et al.,2011).

For this study, factor 9.2 (Lorezo-Seva & Ferrando,2013) was used to conduct factor analysis since it uses tetrachoric correlation. These analyses were performed combining the male and female groups, as well as separately to establish dimensional factorial invariance. This assumption can be approximated by assessing the ratio of first to second eigenvalues, which is an index of the strength of the first dimension of the data (Reise & Waller, 1990). This means that when the first factor explains a large proportion of the total variance, then assumption of unidimensionality has been met.

Table 1. Exploratory factor analysis (EFA) results of subgroups

	Factors	Eigenvalue	Proportion of variance (%)	Cumulative Proportion of variance
Groups Combined	1	6.994	31.79	31.79
	2	1.397	6.35	38.15
	3	1.119	5.09	43.24
male	1	7.068	32.13	32.13
	2	1.448	6.58	38.71
	3	1.319	5.99	44.71
female	1	6.994	31.79	31.79
	2	1.446	6.57	38.37
	3	1.204	5.48	43.84

First three eigenvalues for each groups were given in Table 1. Ratios of first eigenvalues to the second ones indicated that unidimensionality assumption was satisfied for mathematics subtest and for each groups.

Local independence means that after conditioning on ability, examinees' responses to the items on the test are likely to be independent (Hambleton et al, 1991). In general, when the unidimensionality is met, assumption of local independence is said to be met. On the other hand, even assumption of unidimensionality is met, local independence can not be satisfied (Lord, 1980).

Table 2: Results of three different IRT-based DIF methods with item purification

	Lord's Chi-square		Lord's Chi-square with Purification		Raju's method	Area	Raju's Area method With purification		LRT method		LRT method With purification	
Item	Stat.	P-value	Stat.	P-value	Stat.	P-value	Stat.	P-value	Stat.	P-value	Stat.	P-value
m1	0.617	0.734	0.454	0.796	-0.119	0.904	0.156	0.875	0.199	0.655	0.259	0.610
m2	**7.572**	**0.022 ***	**7.913**	**0.019 ***	2.178	**0.029 ***	**3.048**	**0.002****	**4.466**	**0.034 ***	**4.466**	**0.034 ***
m3	1.262	0.532	1.355	0.507	1.085	0.278	1.921	0.054 .	0.690	0.405	0.836	0.360
m4	5.730	0.057 .	6.139	0.046 *	1.651	0.098	2.901	**0.004 ****	0.179	0.671	0.149	0.699
m5	0.325	0.849	0.555	0.757	-0.351	0.725	-1.795	0.072 .	1.067	0.301	1.173	0.278
m6	2.084	0.352	2.476	0.289	0.048	0.961	0.218	0.827	0.122	0.726	0.169	0.680
m7	6.068	0.048 *	6.816	0.033 *	-1.305	0.191	-0.696	0.485	2.817	0.093 .	2.616	0.105
m8	8.602	0.013 *	9.685	0.007 **	-2.449	**0.014***	-3.134	**0.001 ****	3.170	0.075 .	3.042	0.081 .
m9	0.489	0.782	0.635	0.727	0.345	0.730	0.960	0.336	0.153	0.695	0.217	0.640
m10	0.574	0.750	0.641	0.725	-0.748	0.454	0.683	0.494	0.613	0.433	0.505	0.477
m11	3.067	0.215	3.575	0.167	-0.448	0.653	-0.445	0.655	0.012	0.910	-0.012	1.000
m12	**11.09**	**0.003 ****	**11.25**	**0.003 ****	2.547	**0.011 ***	**3.621**	**0.00 *****	**6.243**	**0.012 ***	**6.243**	**0.012 ***
m13	3.551	0.169	3.946	0.139	1.881	0.059 .	4.432	**0.00*****	0.179	0.671	0.223	0.636
m14	6.386	0.041 *	6.912	0.031 *	-0.475	0.634	1.119	0.263	2.394	0.121	2.219	0.136
m15	0.165	0.920	0.334	0.846	-0.084	0.932	-0.250	0.802	0.671	0.412	0.778	0.377
m16	7.587	0.022 *	8.023	0.018 *	2.125	**0.033 ***	3.089	**0.002****	2.004	0.156	2.247	0.133
m17	0.770	0.680	1.005	0.604	0.345	0.729	1.269	0.204	0.148	0.699	0.221	0.638
m18	3.118	0.210	3.605	0.164	0.107	0.914	1.397	0.162	0.176	0.674	0.123	0.725
m19	2.121	0.346	2.313	0.314	-1.216	0.223	-0.802	0.422	11.80	0.0***	11.80	0.00***
m20	0.099	0.951	0.056	0.972	0.310	0.756	-0.474	0.635	0.339	0.560	0.418	0.517
m21	1.240	0.537	1.688	0.429	-1.086	0.277	-2.313	**0.02 ***	-0.022	1.000	-0.03	1.000
m22	3.286	0.193	3.920	0.140	-1.407	0.159	-1.709	0.0874	0.882	0.347	0.800	0.371

Sig. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

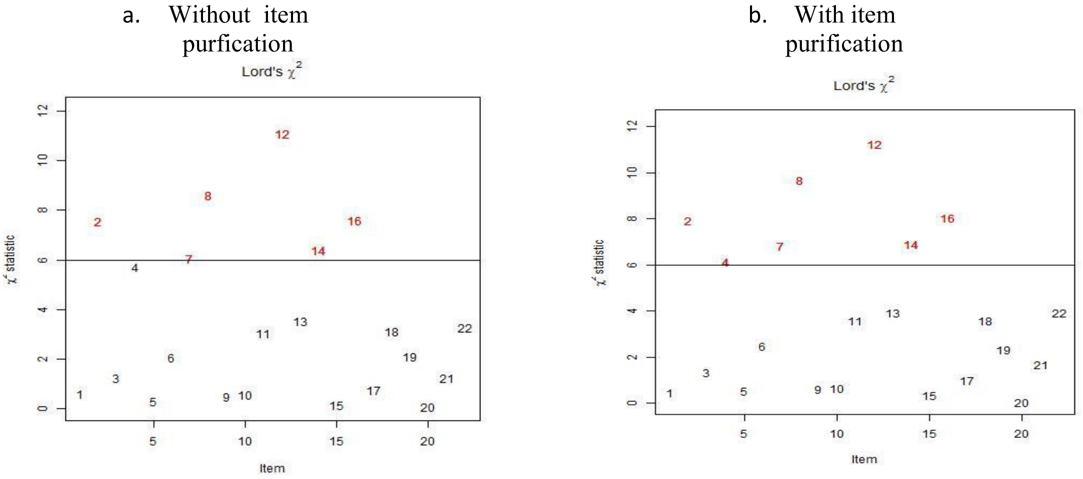
Table 2 shows the DIF statistic and p significance values obtained by Lord's Chi-Square, Raju's Area and Likelihood-Ratio Test methods. In addition, item purification was performed for each methods with 50 iteration and results of each methods with item purification was reported in table 2. In addition, significance level was set to 0.05 for each methods.

First, Lord's Chi-Square method without item purification was used in order to determine DIF items in TIMSS 2011 mathematics subtest booklet 2 and results were reported in table 2. Lord's Chi-Square method results indicates that m2, m7, m8, m12, m14 and m16 items were identified as DIF items and the other 16 items were not detected as DIF items. The fourth column shows the Lord's Chi-Square statistic values obtained in the last step of the purification process, when DIF items are discarded from the computation of sum scores. The corresponding *p* values are also displayed, and the significance levels are indicated with one or more asterisks. This indicates that all items flagged as DIF on the basis of the significance test can be considered to be largely affected by DIF.

Item purification was also performed for Lord's Chi-Square method with purification results indicate that m2, m4, m7, m8, m12, m14 and m16 were detected as functioning differently after 50 iterations and 15 out of 22 items were not detected as DIF. When compared the results of Lord's Chi-Square method with item purification, 6 items detected as DIF were identical and only m4 appeared to show DIF with purification method. They can also be found in table 2 as items with at least one asterisk.

Second, Raju's Area method without item purification was used in order to determine DIF items in TIMSS 2011 mathematics subtest booklet 2 and results were reported in table 2. Raju's Area method results indicates that m2, m8, m12, and m16 items were identified as DIF items and the other 18 items were not detected as DIF items. Item purification was also performed for Raju's Area method and results indicate that m2, m4, m8, m12, m13, m16 and m21 were always classified as DIF items after 50 iterations and 15 out of 22 items were not detected as DIF. When compared the results of Raju's Area method with item purification and without item purification, 4 items detected as DIF were identical and only m4, m13 and m21 appeared to show DIF with purification method.

As third method, Likelihood-Ratio Test (LRT) without item purification was used in order to determine DIF items in TIMSS 2011 mathematics subtest booklet 2 and results were reported in table 2. LRT method results indicates that m2, m12, and m19 items were identified as DIF items and the other 19 items were not detected as DIF items. Item purification was also performed for LRT method and results indicate that m2, m12, and m19 were always classified as DIF items after 50 iterations and 19 out of 22 items were not detected as DIF. When compared the results of LRT method without item purification and with item purification, all three items detected as DIF were identical. Compared to other methods, LRT seems to fail detecting DIF items.

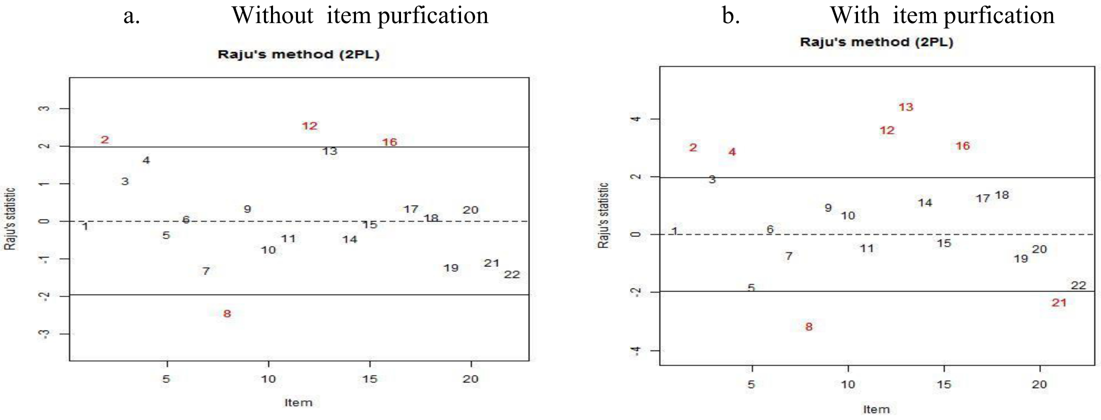


a. Without item purfication

b. With item purification

Figure 1. Lord's Chi-Square statistics and detection threshold with the mathematics data set**.**

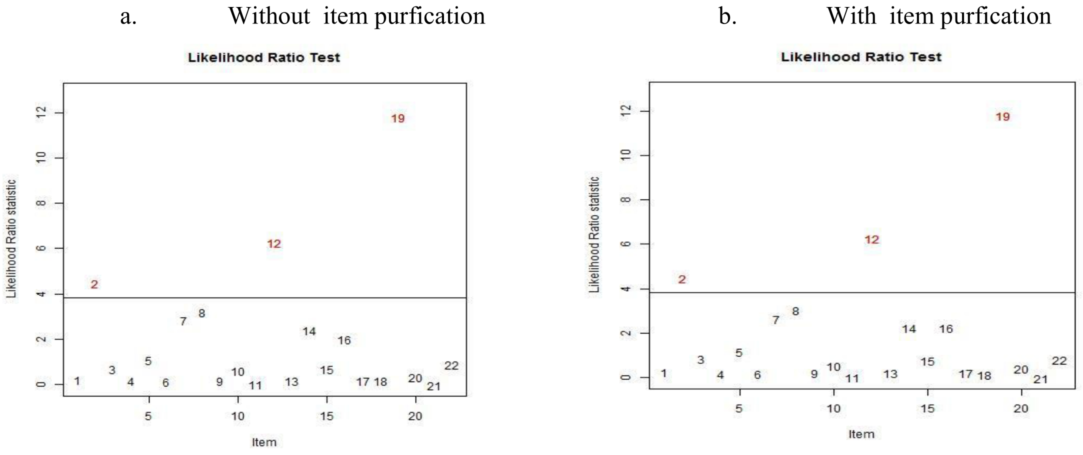
Items are represented by integers referring to their rank in the output list of Figure 1 (1 for the *m1* item, etc.). Both Lord's Chi-Square statistics (Detection threshold: 5.9915, p: 0.05) without item purification and with item purification were presented in Figure 1a and Figure 1b, respectively. Items m2, m7, m8, m12, m14 and m16 were detected as DIF items without item purification. With item purification item m4 was also detected to be DIF item. It can be seen from the table that Item m7is borderline for DIF without item purification, while item 4 is borderline with item purification method. The obtained positive effect size values mean that men are more inclined than women to actually mathematics independent of their degree of inclination to the other lessons in TIMSS 2011.


a. Without item purfication

b. With item purfication

Figure 2. Raju’s Area statistics and detection threshold with the mathematics data set

Raju’s z statistics based on signed area (Detection thresholds: -1.96 and 1.96, p: 0.05) without item purification and with item purification were presented in Figure 2a and Figure 2b, respectively. Items m2, m4, m7, m8, m12, m14 and m16 were detected as DIF items. With item purification item m13 and m21 was also detected to be DIF items while item m7 and m 14 were not functioned differently. The obtained negatif effect size values for item 8 and item 21 mean that unlike other items, these two items were in favor of women rather than men independent of their degree of inclination to the other lessons in TIMSS 2011.


a. Without item purfication

b. With item purfication

Figure 3. LRT statistics and detection threshold with the mathematics data set.

Likelihood Ratio statistics (detection threshold: 3.842, p: 0.05) without item purification and with item purification were presented in Figure 3a and Figure 3b, respectively. With LRT method, items m2, m12 and m19 were detected as DIF items. When compared the results of LRT method wihout item purification and with item purification, all three items detected as DIF were identical. The obtained positive effect size values mean that men are more inclined than women to actually mathematics independent of their degree of inclination to the other lessons in TIMSS 2011.

Table 4: Items Detected As DIF with Three Different IRT-Based Methods

Lord's Chi-square		Raju's Area method		LRT method	
With item purification	Without item purification	With item purification	Without item purification	With item purification	Without item purification
m2	m2	m2	m2	m2	m2
m7	m4	m8	m4	m12	m12
m8	m7	m12	m8	m19	m19
m12	m8	m16	m12		
m14	m12		m13		
m16	m14		m16		
	m16		m21		

Table 4 shows the items detected as DIF by three different IRT-based methods with item purification and without item purification. As can be seen from the table 4, performing item purification tended to increase the number of DIF items except for LRT methods.

5. Results and Discussion

In this study, three different IRT-based DIF methods were used to determine items which functions differently with respect to gender of students in TIMSS 2011 mathematics subtest and results were compared. In addition, item purification was performed for each method in order to see how item purification effected the number of DIF items and DIF statistics compare results of these methods.

Comparing findings from different methods can provide insights into whether differences are due to the different assumptions and criteria embedded within the methods. Moreover, convergent findings across methods are more likely to prompt content experts to modify or remove items with consistent DIF of high magnitude (Yang et al., 2011).

Results indicated that two items (m2, m12) were identified as DIF items by all three methods, whereas 12 other items were never identified as such. For four items (m2, m8, m12 and m16), the Lord's Chi-square and Raju's Area methods identified them as DIF, but the other methods did not. On the other hand, m19 item was detected as DIF item by only LRT methods.

Although, almost all items detected as DIF with three different methods were in favor of male students, Raju's signed area method with item purification indicated that item 8 and item 21 were in favor of female students rather than male students with respect to mathematics subject.

Performing item purification with Lord's Chi-square and Raju's Area methods effected both the number of DIF items and DIF items themselves. However, Performing item purification with LRT method did not affect the number of items detected as DIF.

According to the results, Lord's Chi-square method tended to be more sensitive than other two methods with respect to detecting DIF items. On the other hand, even item purification was performed, LRT method failed to detect many items detected as DIF items by other methods. As it is assumed, these three IRT-based techniques showed substantial agreement in the detection of DIF among the same set of mathematics subtest items, but vary in the number of items flagged with DIF due to different assumptions and criteria used.

This has been a theoretical review of possible IRT-based DIF methods that can be used with a dichotomously scored large scale mathematics test. Although, number of items that displayed DIF differed because of different criteria being used by different methods, it is also important to examine the item carefully in order to try to explain why the item displays DIF.

Finally, Results indicate that there is no single method can be guaranteed to identify all of the DIF items in a test. Not only IRT-based methods but also Non-IRT-based methods should be used to address the instability problem which undermines the utility of current methods and results of both IRT-based and Non-IRT-based methods can be compared in order to determine the best method that detect DIF items accurately.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, **29**, 67-91.

Borsboom, D. (2006).When does measurement invariance matter? *Medical Care*, 44(11) Suppl 3, S176-S181

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement*, **17**, 31-44.

Dorans, N. J., & Holland, P. W. (1993). *DIF detection and description: Mantel-Haenzel and standardization*. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Embretson, S. E. & Reise, S. P., (2000). *Item Response Theory for psychologists.* Mahwah, New Jersey: Lawrence Erlbaum Associates

Ferrando, P.J. & Lorenzo-Seva*, U. (2013).* Unrestricted item factor analysis and some relations with item response theory. *Technical Report. Department of Psychology, Universitat Rovira i Virgili, Tarragona*.

Glas. C. A. W & Verhelst, N. D. (1995) Testing the Rasch Model. *Rasch Models. Foundations, Recent Developments, and Applications.* New York: Springer; 1995. pp. 69–95

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentialy biased test items: Comparison of IRT and Mantel-Haenszel methods. *Aplied Measurement in Education*, 2(4), 313-34.

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

Hambleton, R. K., Clauser, B. E., Mazor, K. M. & Jones, R. W. (1993). Advances in the detection of differentially functioning test items. *European Journal of Psychological Assessment***, 9**, 1-18.

Hambleton, R.K. (2006). Good practices for identifying differential item functioning. *Medical Care*, 44(11) Suppl 3 Suppl 3, S182-S188.

Keser, Ö. F. (2005). Recommendations towards developing educational standards to improve science education in Turkey. *The Turkish Online Journal of Educational Technology – TOJET*, 4(1), Article 6

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

Magis, D., Beland, S., Tuerlinckx, F. & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods* 2010, 42 (3), 847-862

Magis, D., & Facon, B. (2012). Angoff's Delta method revisited: improving the DIF detection under small samples. *British Journal of Mathematical and Statistical Psychology*, *65*, 302-321.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, **17**, 297-334.

Millsap, R.E. (2006). Comments on methods for the investigation of measurement bias in the Mini-Mental State Examination. *Medical Care*, 44(11) Suppl 3 Suppl 3, S171-S175.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd ed.). Thousand Oaks, CA: Sage.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 125-167). Amsterdam: Elsevier.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, **53**, 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, **14**, 197-207.

Reise, S. P. & Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology,* 65, 143– 151

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*. 17(5):1–25

Yang, F. M., Heslin, K. C., Mehta, K. M., Yang, C. W., Ocepek-Welikson, K., Kleinman, M., Morales S. L., et al. (2011). A comparison of item response theory-based methods for examining differential item functioning in object naming test by language of assessment among older Latinos *Psychol Test Assess Model.* 2011 Fall; 53(4): 440–460.

Uzun, S., Butuner, S. Ö., & Yigit, N. (2010). A Comparison of the Results of TIMSS 1999-2007: The Most Successful Five Countries-Turkey Sample. *Elementary Education Online,* 9(3), 1174-1188, 2010.

Wiberg, M. (2007).Measuring And Detecting Differential Item Functioning In Criterion-Referenced Licensing Test A Theoretic Comparison Of Methods EM No 60, ISSN 1103-2685