Eğitim ve Bilim
2009, Cilt 34, Sayı 152

Education and Science
2009, Vol. 34, No 152

A Comparison of the Parametric Methods Based on the Item Response Theory in Determining Differential item and Test Functioning

Değişen Madde ve Test Fonksiyonunun Belirlenmesinde Madde Tepki Kuramı'na Dayalı Parametrik Yöntemlerin Karşılaştırılması[*]

Tuncay ÖĞRETMEN[**]
Ege University

Öz

Bu araştırma, değişen madde ve test fonksiyonlarını belirlemede, Madde Tepki Kuramı'na bağlı parametrik metotların karşılaştırılmasını amaçlamaktadır. Bu amaçla, araştırma, madde parametreleri için karşılaştırma metodu, olabilirlik oranı testine dayalı karşılaştırma metodu, madde ve test düzeyinde değişen madde ve test fonksiyonlarını belirlemede kullanılan metotları kullanarak değişen madde ve test fonksiyonunu analiz etmekte ve elde edilen verilerle bu üç yöntemi karşılaştırmaktadır. Bu çalışma, *Uluslararası Okuma Becerilerinde Gelişim Projesi (PIRLS) 2001* testlerinden elde edilen verilerle yürütülmüştür. Bulgular, bu araştırma kapsamında kullanılan yöntemlere göre elde edilen değişen madde ve test fonksiyonu sonuçları arasında farklılık olduğunu göstermiştir.

Anahtar Sözcükler: Madde Tepki Kuramı, değişen madde ve test fonksiyonu.

Abstract

This study aims to compare parametric methods based on the item response theory in determining differential item and test functioning. To this end, it analyzes differential item and test functioning by using the comparison method for item parameters, the comparison method based on the likelihood ratio test, and the method to determine differential item and test functioning (DFIT) both at the item and test levels, and compares these three methods in terms of the data obtained. The study was conducted on the data on the Progress in International Reading Literacy 2001 (PIRLS-2001) tests. The results of the analyses indicated a differentiation between the results of differential item and test functioning, which were obtained using the methods in question.

Keywords: Item response theory, Differential item and test functioning.

[*] 1. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi'nde sözlü bildiri olarak sunulmuştur.

[**] Yrd. Doç. Dr. Tuncay ÖĞRETMEN, Ege Üniversitesi, Eğitim Fakültesi, tuncay.ogretmen@ege.edu.tr

Introduction

Differential item functioning (DIF) could be defined as the different probability of giving the right answer to a test item by two individuals with the same ability level, but from different groups (Adams & Rowe, 1988; Mellenberg, 1989; Hambleton, Swaminathan and Rogers; 1991; Schrum & Salekin, 2006; Crane, Gibbons, Narasimhalu, Lai & Cella, 2007; McCarty, Oshima and Raju, 2007). In other words, these analyses are conducted to investigate whether educational and psychological measurements of structures differ in terms of groups (Mellenberg, 1983). If a test will be used for a quite heterogeneous universe, then the differential item functioning analyses become the most important part of the item selection process (Crocker & Algina, 1986; Mellenberg, 1983).

In studies on differential item and test functioning or measurement invariance, two major approaches have come into prominence in the relevant literature in recent years. Linear methods known as multisampling confirmatory factor analysis as a part of structural equation modeling fall under the first group, whereas non-linear methods as a part of the item response theory belong to the second group. According to the item response theory, there are three different parametric methods in determining differential item and test functioning, which are a) comparison of item parameters estimated for groups, b) measuring the area between item characteristic curves of the groups, c) a comparison between the probability functions by evaluating the model-data fit for item responses in the groups (Holland & Wainer, 1993; Camili & Shepard, 1994: pp. 46-100; Rodney & Drasgow, 1990; Devine & Raju, 1982; Lord, 1980; Rudner et al., 1980).

On the basis of one-parameter logistic model as part of the item response theory and the partial scoring method, one of the IRT models (Bertrand & Boiteau, 2003, Masters, 1982), this study aims to investigate a) comparison method for item parameters, b) comparison method based on the likelihood ratio test, and c) the detection methods for Differential Item and Test Functioning (DFIT) both at the item and test levels, and to compare the results obtained through these three methods.

Method

The Study Group and the Measurement Instrument

The study was conducted on the data obtained from the reading achievement tests administered to the Turkish and American students who participated in PIRLS-2001 (the Progress in International Reading Literacy-2001) by the International Association for the Evaluation of Educational Achievement (IEA). Within the scope of the study, the data on the MICE reading tests – one of the reading literacy tests of PIRLS – were used. The MICE test consists of 14 questions. 7 of these (items 1, 2, 3, 5, 8, 9 and 13) are multiple-choice questions scored as 1-0, and 5 (items 4, 7, 10, 11 and 14) are items requiring short answers again scored as 1-0, whereas the 6th and 12th items require long answers again, the first of which was scored as 0-1-2 and the second as 0-1-2-3 through partial scoring. Table 1 presents the distribution of the items in the MICE reading test according to the processes of comprehension they are assumed to measure (Gonzalez & Kennedy, 2003; Mullis, Martin, Gonzalez and Kennedy, 2003).

Table 1.
Distribution of the Items in the MICE test according to the Processes of Comprehension Assumed to Measure

Reading Process	Items
Focus on retrieve explicitly stated information and ideas	2, 5, 10
Making straightforward inferences	1, 3, 7, 9
Interpret and integrate ideas and information	4, 6, 11, 12
Examine and evaluate content, language, and textual elements	8, 13, 14

The data required for the study were obtained from the website of the ISC (International Study Center), 2003.

Data Analysis

To find answers to the study problem, it was examined whether the application of the MICE reading test and its items as part of the PIRLS project in Turkey and the USA display differential item and test functioning across cultures using the parameter comparison method, likelihood ratio test model comparison method and detection method for differential item and test functioning (DFIT) both at the item and test levels. These methods are briefly described as follows.

Comparison Method for Item Parameters in Determining Differential Item Functioning:

To measure differential item functioning at the item level, parameter difference statistics for item discrimination and item difficulty are first calculated for each item from the reference and focal groups using the following equations (Reise, Smith and Furr, 2001; Smith, 2002).

Differential item parameter difference = $\hat{a}_{i\,(R)}$- $\hat{a}_{i(F)}$ (1)

Differential item parameter difference = $b_{i\,(R)}$- $b_{i(F)}$ (2)

These equations yield a direct measurement of estimated item value difference for a particular item by subtracting the estimated parameter value for the focal group from either the item discrimination or the estimated value of item difficulty for the reference group. In the next stage, these difference values obtained from the parameters are standardized. For instance, the standardized differential item functioning (SDIF) value is obtained through the following equation (Reise, Smith and Furr, 2001; Smith, 2002).

$$\text{SDIF} = \frac{DIF}{\sqrt{\operatorname{var}\hat{b}_{i(R)} + \operatorname{var}\hat{b}_{i(F)}}} \quad (3)$$

Difference statistics of the standardized differential item functioning are similar to standard scores and yield a measurement of the contrast between the estimated item parameters for the compared groups. The square of the standardized difference value could be evaluated as χ^2 statistics under 1 degree of freedom (du Toit, 2003). Thus, if an item has a significant χ^2 value at the 0.01 or 0.05 alpha level of significance, then it is considered to display differential item functioning across groups.

The Model Comparison Method Based On the Likelihood Ratio Test in Determining Differential Item Functioning:

Another method of examining DIF for both dichotomous and polytomous response models is the model comparison method, which is based on the significance of the likelihood ratio difference between two models described as compact and augmented models (Thissen, Steinberg and Wainer, 1993; Sireci & Berberoğlu, 2000). The method is illustrated in equation 18.

Likelihood Ratio (LR)= L*(CompactModel) / L* (AugmentedModel)(4)

The augmented model incorporates one or more parameters freed to be estimated. For instance, parameters *a* and *b* are typically allowed to differ from one group to another for the tested items. The degrees of freedom are calculated as the difference between the number of parameters in the augmented and compact model. The resulting G^2 statistic is distributed as chi-square, and if this value is greater than the chi-square table value, it is assumed that there are considerable differences in item parameter differences between the two groups. If the result is not significant, it is concluded that none of the parameters of the augmented model is different than 0 (Kim, Cohen, DiStefano and Kim, 1998; Meade & Lautenschlager, 2004; Teresi, Kleinman, Welikson, 2000). The aim of this model comparison is to test whether there is really a need for the parameters added to the augmented model and whether they differ significantly from zero in improving the model's goodness of fit. Here, assuming that the compact model of the H_0 hypothesis includes only N number of item parameters, whereas the augmented model of the H_1 hypothesis incorporates M number of parameters in addition to the N parameters, the compact model as a simpler one will have less number of parameters compared to the augmented model.

The likelihood test seeks an answer to the question of whether the data sampling confirms the H_0 hypothesis. In other words, do the M number of added parameters improve the model's goodness of fit? If the likelihood ratio is log transformed -2 times, then the result will be a test statistic displaying a χ^2 distribution with M degrees of freedom particularly in large size samplings.

χ^2 (M) ≈ -2ln(LR) = [-2lnL*(Compact Model)]- [-2lnL*(Augmented Model)] (5)

This result is calculated through the mathematical equation *ln(x/y) = lnx-lny*. Model comparison method for likelihood test involves the subtraction of the values calculated from the -2 times logarithmic transformation of the compact and augmented models.

Prior to the DIS analysis performed by the likelihood test, a compact model is formulated assuming that the examined items do not include the DIF. In the compact model, scaling is performed by equating the item parameters for both groups. In the second stage, an augmented model is formulated in parallel with the number of items investigated through DIF. Assuming that a particular item includes DIF in each model, the parameters of that item are freed in the reference and focal groups, and the item parameters of other items are equated. Thus, k+1 number of analyses are performed, once for the compact model, and k times for the augmented model (Meade & Lautenschlager, 2004; Holland & Wainer, 1993).

While examining differential item functioning through the likelihood analysis, anchor items consisting of the items of the test itself are needed as internal criteria to equate the group parameters on the same metric (Thissen, Steinberg and Wainer, 1993). Here, the probabilities of the compact model, which assumes that item parameters do not differ across groups, are compared to those of the augmented model, in which the items other than the anchor items are examined for the differential item functioning. The anchor items are assumed as items that do not display differential item functioning and thus, they are not subjected to differential item functioning analysis (Kim & Cohen, 1995).

In order to select the anchor items for the analyzed tests, this study uses the iterative item selection process proposed by Kim and Cohen (1995). The steps of this process are as follows:

1. The -2 logL (G_1^2) value for the compact model is calculated where all item parameters are equated across the compared groups.
2. The -2 logL (G_2^2) values for the augmented models are calculated, where the parameters for each item are freed in the compared groups.
3. The -2 logL values calculated from the augmented model in which the parameters are freed for each item in the test are subtracted from the -2 logL value calculated from the compact model, in which all item parameters were initially equated across the groups (where the parameters were constant in the compared groups). That is, the test statistic values of $G_i^2 = G_{compact}^2 - G_{augmented}^2$ are found. The obtained value is then compared to the χ^2 table value at the relevant degree of freedom, and if any significant G_i^2 value is not found, the operation is stopped at this stage. If there are items with significant G_i^2 values, this test item with the greatest value is removed from the test.
4. A compact model is re-established for the remaining items and the -2 logL is obtained.
5. New augmented models are formulated in parallel with the number of remaining items, and their -2 logL values are calculated.
6. The test statistic values of $G_i^2 = G_{compact}^2 - G_{augmented}^2$ are once more calculated and only the item yielding the greatest G_i^2 value is removed from the test. This iterative item-sorting process is continued until the point where none of the remaining items of the test yield a significant G_i^2 value. Consequently, a set of anchor items is established through these items which are estimated to display no differential item functioning.

The likelihood ratio test and differential item functioning analyses continue with an augmented model, where only the parameters of anchor items are equated across the compared groups and other items likely to display differential item functioning are freed across the groups; and an analysis is made to calculate the -2 logL value is for this augmented model. Subsequently, compact models are formulated in which the parameters of each item other than the selected anchor items likely to display differential item functioning are constrained (equated) across the compared groups. The G_i^2 values are calculated by subtracting the -2 logL value obtained for each item in the compact model from the -2 logL value of the augmented model in which only the anchor items are equated. The obtained G_i^2 value is compared to the χ^2 table value at the relevant degrees of freedom, and the analyses are finalized by concluding that items yielding a G_i^2 value greater than the table value display differential item functioning.

DFIT Method in Determining the Differential Item and Test Functioning

Raju, van der Linden, and Fleer (1995) suggested a very useful method determining the differential item and test functioning (DFIT) at the item and test level. The method offers two basic approaches. The first is the two index calculation methods, which are the Differential Test Functioning (DTF) Index and the compensatory or signed differential item functioning index (CDIF) for each item in the test. As the total of the CDIF values calculated for all the test items will be equal to the differential test functioning value, iteration operations will be repeated until a statistically insignificant test functioning value is reached removing the items with greater and negative CDIF value at each iteration. This process is explained as given in the following formulae:

$$DTF = \varepsilon_j(D^2_j) = \sigma^2_{Dj} + \overline{D^2_j} \qquad (6)$$

$$DTF = \Sigma_i\, CDIF_i \qquad (7)$$

$$CDIF_i = \varepsilon_j\,(d_{ij}D_j) = \sigma_{dijDj} + \overline{d_{ij}\, D_j} \qquad (8)$$

where $d_{ij} = P_{iR}(\theta_j) - P_{iF}(\theta_j)$ and $D_j = V_R(\theta_j) - V_F(\theta_j)$ at the $\boldsymbol{\theta}_j$ level for item **i**.

The second approach incorporates the non-compensatory or "unsigned" DIF Index (NCDIF), which is also referred as the area index giving the value of area between Item Characteristic Curves (ICC) which are obtained from sub-populations (Betrand and Boietau, 2003).

NCDIF is formulated through the following equation:

$$NCDIF_i = \varepsilon_j\,(d^2_{ij}) = \sigma^2_{dj} + \overline{d^2_j} \text{ where } d_{ij} = [P_{iR}(\theta_j) - P_{iF}(\theta_j)] \text{ and } j= 1, 2, \ldots, \mathbf{n_F}\,. \qquad (9)$$

and $\mathbf{n_F}$ is the number of respondents in the focal group.

At $\mathbf{n_F}$ degrees of freedom, the statistic regarding the chi-square test is as follows:

$$\chi^2 = n_F{*}\ NCDIF_i\, /\ \sigma^2_{dij} \qquad (10)$$

As it is unsigned, the NCDIF index will always have a positive value. In case the NCDIF value obtained for the items is greater than 0.006 and the relevant chi-square value is significant at the significance level of α= 0.01, it is concluded that the items display differential item functioning (McCarty, Oshima and Raju, 2002; Raju, van der Linden and Fleer, 1995). In favor or at the expense of which group the items display differential item functioning will be determined by looking at the sign of the CDIF index.

In the analyses conducted through each of the three methods, it was examined whether only parameter *b* (item difficulty) displayed differential item functioning across the groups. During the analyses carried out according to each of the three methods, the "*a*" value of item discrimination was equated to 1 for all items. Thus, since the items test items were scored by two or more scoring categories, one-parameter logistic item response model was used for the *Generalized Partial Credit Model* (Muraki, 1992) in the analyses of each of the three methods.

Under the scope of this study, the computer program PARSCALE 4.1 was used for the differential item functioning analyses based on the comparison method for the item parameters (Muraki and Bock, 1996). The likelihood ratio test and the differential item functioning analyses based on the model comparison method were performed using the

MULTILOG 7.03 program (Thissen, 1992). The computer program DFITD6 was used in the differential item and test functioning analyses (Raju, 2004). Furthermore, the EQUATE99 program was utilized to perform the metric equation operations for the DFIT procedures (Stark, 1999).

Results and Interpretations

Results of the DIF Analysis Performed Through the Parameter Comparison Method:

Results of the DIF analysis for the MICE reading passage performed on the American and Turkish samplings through the parameter comparison method are provided in Table 2.

The DIF analysis performed on the American and Turkish samplings through the parameter comparison method demonstrates that items 1, 4, 6, 8, 10, 12, 13 and 14 display differential item functioning. Among these, items 8, 10, 12 and 14 display DIF at the significance level of 0.01. A comparison between the response rates for scoring categories in terms of the countries in question revealed that all the items displaying DIF are in favor of the American sampling. For instance, item 6 is an item partially scored as 0-1-2. While the ranking for this item in terms of scoring categories was observed to be 12%, 24.7% and 63.3% in the American sampling, the same ranking turned out to be as 36%, 16.6% and 47.4% for the Turkish sampling. In the case of item 8, which was rated under two categories, while the rate of correct answers in the American sampling was 79.5%, the rate of those giving the correct answer in the Turkish sampling was observed as only 38.6%.

Table 2.
Results of the DIF Analysis Performed through the Parameter Comparison Method in the American and Turkish Samplings for the Assessment Questions on the MICE reading passage

Item No	Item Difficulty (b_i)		Contrast	χ^2 (sd=1)	p
	American Samplings	Turkish Samplings			
1 (0-1)	-1,36	-1,71	-0,35	4,89*	0,02
2 (0-1)	-1,97	-1,80	0,17	0,71	0,40
3 (0-1)	-0,90	-0,85	0,04	0,11	0,73
4 (0-1)	0,56	0,34	0,29	4,17*	0,03
5 (0-1)	-1,44	-1,48	-0,40	0,06	0,79
6 (0,1,2)	-0,82	-1,10	-0,28	6,02*	0,01
7 (0-1)	-1,05	-1,23	-0,17	1,48	0,22
8 (0-1)	-0,38	0,30	-0,68	24,48**	0,00
9 (0-1)	-1,31	-1,30	0,01	0,00	0,90
10 (0-1)	-1,41	-2,07	-0,66	17,22**	0,00
11 (0-1)	-0,03	0,02	0,57	0,16	0,68
12 (0,1,2,3)	0,07	-0,45	-0,53	27,35**	0,00
13 (0,1)	-1,16	-0,88	0,28	3,87*	0,04
14 (0,1)	-1,11	-0,61	0,14	12,12**	0,00
TOPLAM				102,02**	0,00

*p<0,05 ; **p<0,01

Note: The values in paranthesis indicate the scoring pattern for the items.

It has been showed that, in giving the answers requiring high θ level of skill, students of the Turkish sampling experience more difficulties when compared to those of the American sampling at all items displaying DIF.

Results of the DIF Analysis Performed Through Model Comparison Method Based on the Likelihood Ratio Test:

As a result of the Ist iterative item selection process performed for the selection of the anchor items in the differential item functioning analysis for the MICE reading passage assessment test, the difference values between the compact and augmented models of items 10, 11 and 13 were found to be significant at the significance level of 0.05 and those of items 4, 6, 8 and 13 at the significance level of 0.01.

After removing item 8, which yielded the greatest G_i^2 difference value ($\chi^2_{(4;\ 0.01)}$= 55.2, p<0.01), the analysis proceeded with the IInd selective iteration process for the remaining 13 items. The -2logL value of the compact model was obtained as 2270.2 in the IInd iteration. Then removing item 14, which yielded the greatest G_i^2 difference value, we moved on to the IIIrd selective iteration process for the remaining 12 items. In the IIIrd iteration, the -2logL value of compact model was calculated as 1563.1. As the likelihood ratio value of the compact model which was established for the IVth iteration turned out to be negative, the analyses could not be finalized. Table 3. presents information regarding the three iterations through which the iterative item selection processes were performed.

Table 3.

Comparison Results for the MICE reading passage assessment test at the I., II. and III. Iterative Item Selection Processes in terms of Compact and Augmented Models

Item no	I. Iteration $G_1^2 - G_2^2$	II. Iteration $G_1^2 - G_2^2$	III. Iteration $G_1^2 - G_2^2$
1	4	0,8	0,2
2	6,4	4,1	6,3
3	7,7	6,2	8,5
4	17,6**	18,6**	23,8**
5	3,9	1,5	3,1
6	23,8**	20,3**	21,4**
7	2,9	0	0,4
8	55,2**	Çıkarıldı	Çıkarıldı
9	5,7	3,9	6,6
10	12,4*	8,9	5,6
11	10,4*	10,1*	13,3**
12	41,1	31,6**	23,2**
13	19,6*	8,6	9,4
14	43,3**	44,8**	Çıkarıldı

$\chi^2_{(0,01)}$= 13,28, sd.=4, **p<0,01

$\chi^2_{(0,05)}$= 9,49, sd.=4, *p<0,05

As seen in Table 3, upon the three iterative item selection processes, items 3, 5, 9 and 13, whose difference of fit for the compact and augmented models was found to be statistically insignificant in each of the three iterations, were identified as the anchor items since they do not tend to display differential item functioning.

Having completed the process of selecting the anchor items, item parameter estimations and the -2logL values for the compact and augmented models were re-calculated. Thus, in the augmented model established to this end, only the parameters of items 3, 5, 9 and 13 were equated in the reference and focal groups, which was not applied to parameter values of other items across the groups. Subsequently, compact models were established, where each of the items other than the anchor items were equated to the anchor items. The analysis continued with a comparison between the likelihood ratios of the compact and augmented models, where likelihood ratio value of the former constraining the item was subtracted from that of the latter in which it was freed.

Table 4.

2logL Values and the Differential Item Functioning Analysis Results for the Compact and Augmented Models of MICE Reading Text

Item no	Compact model G_1^2	Augmented model G_2^2	$G_1^2 = G_1^2 - G_2^2$
1	2925,0	2924,3	0,7
2	2946,3	2924,3	22**
3	Anchor item	Anchor item	Anchor item
4	2947,8	2924,3	23,5**
5	Anchor item	Anchor item	Anchor item
6	2946,2	2924,3	21,9**
7	2928,8	2924,3	4,5
8	3008,9	2924,3	84,6**
9	Anchor item	Bağ maddesi	Anchor item
10	2927,0	2924,3	2,7
11	2942,2	2924,3	17,9**
12	2937,5	2924,3	13,2*
13	Anchor item	Anchor item	Anchor item
14	2976,0	2924,3	51,7**

$\chi^2_{(0.01)}$= 13.28, sd.=4, **p<0.01

$\chi^2_{(0.05)}$= 9.49, sd.=4, *p<0.05

Making such a comparison between the compact and augmented models, it was demonstrated whether the difference between the fit values of the two models is statistically significant. Table 4 presents the -2logL values and the differential item functioning analysis results for the compact and augmented models of MICE reading passage assessment.

The -2logL values of the augmented model in which items 3, 5, 9 and 13 were equated across the groups for the differential item functioning analysis were calculated as 2924.3.

As the results of the analysis suggest, for the the non-anchor items in the MICE reading passage assessment test, the difference between the -2logL values for the compact and augmented models of item 12 was found to be significant at the level of 0.05, and -2logL values for the compact and augmented models of items 2, 4, 6, 8, 11 and 14 were significant at the level of 0.01. This result signifies that there was observed a differential item functioning in intercultural applications of the translated and adapted MICE reading passage assessment test.

Results of the DIF Analysis Performed Through the Differential Item and Test Functioning (DFIT) Method:

Table 5 exhibits the results of the DIF analysis conducted on the American and Turkish samplings for the MICE reading passage assessment test. The results of the analysis demonstrates that items other than items 1, 10 and 12 display differential item functioning in terms of the NCDIF index values. The differential item functioning value was found to be 2.95, which was also found statistically significant at the test level at the cut point level of 0.336 (χ^2 = 5895.10 , sd= 307, p< 0.001).

Table 5.
Results of Differential Item and Test Functioning (DFIT)

Items	CDIF	NCDIF	χ^2	p
1	0,121	0,005	1732,51	0,00
2	0,215	0,017	1555,59	0,00
3	0,247	0,021**	5041,96	0,00
4	0,272	0,028**	8118,44	0,00
5	0,200	0,014	2261,81	0,00
6	0,168	0,010	3521,01	0,00
7	0,167	0,010	2983,85	0,00
8	0,355	0,045**	10228,3	0,00
9	0,228	0,018**	2867,97	0,00
10	0,067	0,002	1138,13	0,00
11	0,227	0,019	9438,1	0,00
12	0,008	0,000	329,27	0,18
13	0,255	0,022**	3692,8	0,00
14	0,421	0,060**	6629,75	0,00

In other words, the differential test functioning was observed across the groups at the test level. It is further understood that if items 3, 4, 8, 9, 13 and 14 are respectively removed from the test, the differential test functioning index value will fall down to 0.312; and taking into consideration the remaining items, no differential functioning will be observed at the test and item level. Therefore, it is now evident that items 3, 4, 8, 9, 13 and 14 are items displaying differential item functioning, and thereby the test also displays differential item functioning across the groups. Looking at the signs of the CDIF index values, it can be observed that all of the items, which display differential item functioning, operate in favor of the American sampling group.

Conclusion and Suggestions

In the MICE reading passage assessment test, it was determined that 8 of the items display differential item functioning according to the parameter comparison method, 7 according to the comparison method based on the likelihood ratio test, and 6 according to the DFIT method. Items 4, 8 and 14 are common for each of the three methods. On the other hand, it is another striking fact that five items were identified as common between the methods of parameter comparison and model comparison based on likelihood ratio test. Taking into account that the processes of comprehension measured by the items of the MICE test, it is observed that the items displaying differential item functioning are rather on the processes of "interpreting and integrating ideas and information" and "examining and evaluating content, language, and textual elements".

During the iterative item procedures performed to select the anchor items in the model comparison method based on the likelihood ratio test, since augmented models are established in parallel with the number of items in each iteration, data files and remaining items are re-defined for each new model, commands required for the software have to be rewritten; the procedures in question require time-consuming and laborious efforts, and give rise to risk of making errors in the transition across models and iterations. In addition to this, using only one command file in the parameter comparison method, it is possible to attain results in a shorter time. Furthermore, the software used for this method is appropriate for multiple category items scored in different ways, and it provides the researcher with several opportunities such as working on as many item parameters as s/he likes, convenience in constraining any parameter s/he likes across the compared groups, testing the difference in the differential item parameters through the χ^2 test and an easier interpretation of the results. Therefore, using the parameter comparison method particularly in DIF analyses for tests with many items could offer facilities to the researcher. However, the DFIT procedures used in determining differential item and test functioning not only provides information on the item level, but also offers statistical proofs about whether the test has differential functioning. Furthermore, this method allows us to easily decide the removal of which items from the test will result in a case where the test might cease to display differential item functioning. The direction of the DIF could further be determined by examining the signed indices. Transforming the parameters of the focal group to the metric of the reference group, and thereby determining the differential item and test functioning is another advantage of the DFIT procedures.

It can be suggested that using the comparison method based on the likelihood ratio test and the parameter comparison method would be more appropriate, whereas DFIT procedures would be rather useful in determining the differential item and test functioning in tests serving the purposes of selection and placement (Korkmaz, 2005; Öğretmen, 2006).

The scope of this study was confined to the item difficulty index 'b_i' in the differential item and test functioning analyses. However, it is also possible to investigate whether item discrimination 'a_i' and guessing parameters 'c_i' display differential functioning across groups and to compare the results in terms of the methods. DIF and test functioning determining methods could also be comparatively studied in terms of situations with high rate of missing data and different sampling sizes.

References

Adams, R. J., & Rowe, K. J., (1988). *Item Bias* in Keeves, J.P. (ed.) Educational research, methodology, and measurement: An international handbook. Oxford: Pergamon Press.

Bertrand, R., & Boiteau, N. (2003). Comparing the Stability of IRT-Based and Non IRT-Based DIF Methods in Different Cultural Context Using TIMSS Data. *ERIC Report-Research (143). EDRS. ED 476 924*. 20p.

Crane, P.K., Gibbons, L.E., Narasimhalu, K., Lai, J. S., & Cella, D. (2007). Rapid detection of differential item functioning in assessments of health-related quality of life: *The Functional Assessment of Cancer Therapy. Quality of Life Research*, 16, 101–114.

Crocker, L., & Algina, J., (1986) *Introduction to Classical and Modern Test Theory.* Orlando: Harcourt Brace Jovanovich.

Devine, P. J., & Raju N. S., (1982) Extent of Overlap Among Four Item Bias Methods. *Educational and Psychological Measurement*, 42, 1049-1066.

Gonzalez, E. J., & Kennedy, A. M. (2003*). "PIRLS 2001 User Guide for the International Database"***.** (IEA) International Study Center, Boston College.

Hambleton, R K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. London: Sage Publication.

Holland, P. W., & Wainer, H. (1993). *Differential Item Functioning*. Hillside, NJ: Lawrence Erlbaum.

IEA (International Association for the Evaluation of Educational Achievement). (n.d.). TIMSS 1999 Publications. Retrieved November 15, 2001 from *http//isc.bc.edu/timss1999i/database.html*

Kim, S. H., & Cohen, A. S. (1995). A Comparison of Lord's Chi-square, Raju's Area Measures, and the Likelihood ratio Test on Detection of Differential Item Functioning. *Applied Measurement in Education*, 8(4), 291-312.

Kim, S. H., Cohen, A. S., DiStefano, C. A., & Kim, S. (1998). An Investigation of the Likelihood Ratio Test for Detection of Differential Item Functioning Under the Partial Credit Model. *ERIC Reports- Evaluative. EDRS. ED 442 837*. 23 p.

Korkmaz, M. (2005). *Madde Cevap Kuramı'na Dayalı Olarak Çok Kategorili Maddelerde Madde ve Test Yanlılığının (İşlevsel Farklılığın) İncelenmesi.* Unpublished PhD Thesis. Ege University, Department of Psychology.

Lord, F. M. (1980). *Applications of Item ResponseTtheory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum.

Master, G. N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrica*, 47, 149-174.

Mc Carty, F. A., Oshima, T. C., & Raju, N.S. (2002). Identifying possible sources of differential functioning using differential bundle functioning with polytomous scored data. Peper presented at the annual meeting of the Amreican Educational Research Association, New Orleans.

McCarty, F. A., Oshima, T. C., & Raju, N.S.(2007). Identifying Possible Sources of Differential Functioning Using Differential Bundle Functioning With Polytomously Scored Data. *Applied Measurement in Education*, *20*(2), 205–225

Meade, A. W., & Lautenschlager, G.J. (2004). A Comparison of Item Response Theory and Confirmatory Factor analytic Methodologies for Establishing Measurement Equivalence/Invariance. *Organizational Research Methods*, 7(4), 361-388.

Mellenberg, G. J. (1983). Conditional Item Bias Methods. *In S. H. Irvine and W. J. Barry (Eds), Human Assesment and Cultural Factors* (pp. 293-302). Newyork: Plenum Press.

Mullis, I.V.S., Martin, M.O., Gonzalez, E. J. & Kennedy, A.M., (2003) *"PIRLS 2001 International Report." (IEA) International Study Center, Boston College.*

Muraki, E. (1992). A Generalized Partial Credit Model: Applications of an EM Algoritm. *Applied Psychological Measurement*, 16, 159-176.

Muraki, E., & Bock, R. D. (1996). *PARSCALE (V4.1). Parameter Scaling of Rating Data.* Chicago, IL: Scientific Software, Inc.

Öğretmen, T. (2006). *Uluslararası Okuma Becerilerinde Gelişim Projesi (PIRLS) 2001 Testi'nin psikometrik özelliklerinin incelenmesi: Türkiye-Amerika Birleşik Devletleri örneği.* Unpublished PhD Thesis, Hacettepe University, Department of Educational Sciences.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement,* 19(4), 353-368.

Raju, N. S. (2004). *DFITP6. A FORTRAN program for calculating DIF/DTF [Computer Software]*. Chicago: Illinois Institute of Technology.

Reise, S. P., Smith, L., & Furr, R.M. (2001). Invariance on the PI-R Neuroticism Scale. *Multivariate Behavioral Research*, 36(1), 83-110.

Rodney, G. L. and Drasgow, F., (1990). Evaluation of Two Methods for Estimating Item Response Theory Parameters When Assessing Differential Item Functioning. *Journal of Applied Psychology.* 75(2), 164-174.

Rudner, L., Getson, P. R., & Knight, D. L. (1980). Biased Item Detection Techniques. *Journal of Educational Statistics*, 5, 213-233.

Stark, S. (1999). *EQUATE99. Computer programme for equatimg two metrics in item response theory.* University of Illinois IRT Laboratory.

Smith, L. L. (2002). On the Usefulness of Item Bias Analysis to Personality Psychology. *Personalityy and Social Psychology Bulletin*, 28(6), 754-763.

Schrum, C. L., & Salekin R.T. (2006). Psychopathy in Adolescent Female Offenders: An Item Response Theory Analysis of the Psychopathy Checklist: Youth Version. *Behavioral Sciences and the Law Behav. Sci. Law.* 24, 39–63.

Teresi, J. A., Kleinman, M., & Welikson, O. K. (2000). Modern Psychometric Methods for Detection of Differential Item Functioning: Application to Cognitive Assessment Measures. *Statistics in Medicine*, 19, 1651-1683.

Thissen, D. (1992). *MULTILOG: Multiple Category Item Analysis and Test Scoring Using Item response Theory* (V7.03). Chiago: Scientific Software International, Inc.

Thissen, D., Steinber, L., & Wainer, H. (1993). *Detection of Differential Item Functioning Using the Parameters of Item Response Model. İç. P.W. Holland and H. Wainer (Ed). Differential Item Functioning* (67-113). Hillside, NJ: Erlbaum.

Toit, M. (2003). IRT *from SSI: Bilog-MG, Multilog, Parscale, Testfact.* Scientific Software International, Inc.