# COMPARISON OF IRT LIKELIHOOD RATIO TEST AND LOGISTIC REGRESSION DIF DETECTION PROCEDURES*

# MTK OLABİLİRLİK ORANI TESTİ VE LOJİSTİK REGRESYON DEĞİŞEN MADDE FONKSİYONU BELİRLEME YÖNTEMLERİNİN KARŞILAŞTIRILMASI

Burcu ATAR**, Akihito KAMATA***

**ABSTRACT**: The Type I error rates and the power of IRT likelihood ratio test and cumulative logit ordinal logistic regression procedures in detecting differential item functioning (DIF) for polytomously scored items were investigated in this Monte Carlo simulation study. For this purpose, 54 simulation conditions (combinations of 3 sample sizes, 2 sample size ratios, 3 DIF magnitudes, and 3 DIF conditions) were generated and each simulation condition was replicated 200 times. In general, the Type I error rates of IRT likelihood ratio test and ordinal logistic regression procedures were in good control across all simulation conditions. The power of likelihood-ratio test was high for medium or large sample sizes and moderate or large DIF magnitude conditions. The power of this procedure increased as the sample size or DIF magnitude increased. On the other hand, the power of ordinal logistic regression procedure was unacceptably low for all DIF conditions except for the large sample size and large DIF magnitude condition.

**Keywords:** differential item functioning, IRT likelihood ratio test, ordinal logistic regression

**ÖZET:** Bu Monte Carlo simülasyon çalışmasında, MTK olabilirlik oranı testi ve kümülatif lojit ordinal lojistik regresyon yöntemlerinin çok kategorili puanlanan maddeler için değişen madde fonksiyonunu (DMF) saptamada tip I hata oranları ve güçleri incelenmiştir. Bu amaç doğrultusunda, 54 simülasyon koşulu (3 örneklem büyüklüğü, 2 örneklem büyüklüğü oranı, 3 DMF büyüklüğü ve 3 DMF durumu) üretilmiş ve herbir simülasyon koşulu 200 kere tekrar edilmiştir. MTK olabilirlik oranı testi ve ordinal lojistik regresyon yöntemlerinin tip I hata oranları genel olarak bütün simülasyon koşulları altında iyi kontrol sağlamıştır. MTK olabilirlik oranı testinin gücü orta veya büyük örneklem büyüklüğü ve orta veya büyük DMF büyüklüğü için yüksek bulunmuştur. Bu yöntemin gücü örneklem büyüklüğü veya DMF büyüklüğü arttıkça artmıştır. Diğer yandan, ordinal lojistik regresyon yönteminin gücü büyük örneklem büyüklüğü ve büyük DMF koşulu hariç bütün simulasyon koşulları için kabul edilemez derecede düşük çıkmıştır.

**Anahtar sözcükler:** değişen madde fonksiyonu, MTK olabilirlik oranı testi, ordinal lojistik regresyon

## 1. INTRODUCTION

It is common to see educational tests that contain only polytomously scored items or educational tests that contain both dichotomously and polytomously scored items. Therefore, evaluating polytomously scored items that exhibit differential item functioning (DIF) is as essential as evaluating dichotomously scored items that exhibit DIF. Under item response theory for polytomosly scored items, it is said that an item functions differentially between groups of examinees with same ability levels when the expected score functions in those groups are not equal (Cohen, Kim, & Baker 1993; Kim & Cohen 1998).

It is important to identify items that function differentially between groups of examinees with same ability levels since those items are threat to the validity of interpretation and use of the test scores. The issues of construct-related evidence of validity and the issues of DIF are interrelated in the sense that the number of constructs being measured by the test or the item (Ackerman 1992). If a test lacks construct-related evidence of validity, it means that the test contains items that are measuring constructs other than those are intended to be measured, indicating that there is a potential for bias against or for a certain group of examinees.

Several procedures are available to detect differential item functioning (DIF) for dichotomously scored items. Commonly used DIF procedures for dichotomously scored items include the Mantel-

[*] Bu çalışma Prof. Dr. Akihito Kamata danışmanlığında yürütülen Burcu Atar'ın doktora tezinden üretilmiştir.
[**] Dr. Burcu Atar, Hacettepe Üniversitesi, e-posta: burcua@hacettepe.edu.tr
[***] Prof. Dr. Akihito Kamata, University of Oregon, e-posta: kamata@uoregon.edu

Haenszel (MH: Holland & Thayer 1988), standardization (Dorans & Kulick 1986), logistic regression (Swaminathan & Rogers 1990), simultaneous item bias test (SIBTEST: Shealy & Stout 1993) procedures and the procedures based on item response theory (Thissen, Steinberg, & Gerard 1986; Thissen, Steinberg, & Wainer 1988, 1993). Many of DIF procedures that are conducted for dichotomously scored items are extended for polytomously scored items. The generalized Mantel-Haenszel procedure (GMH: Zwick, Donoghue, & Grima 1993a, 1993b) as the extension of the MH procedure for nominal data, the standardized mean difference procedure (SMD: Dorans & Schmitt 1991) as the extension of the standardization procedure, the polytomous logistic regression (French & Miller 1996), the logistic discriminant function analysis (Miller & Spray 1993), and the ordinal logistic regression (Zumbo 1999) procedures as the extensions of the logistic regression procedure; and the extension of the SIBTEST procedure (poly-SIBTEST: Chang, Mazzeo, & Roussos 1996) are the examples of the polytomous DIF procedures.

In this study, the IRT likelihood-ratio test and the cumulative logit ordinal logistic regression DIF detection procedures were investigated, as there is not enough research on the effectiveness of these procedures on detecting DIF for polytomously scored items. These two procedures will be described in the following two sections to provide general information about the procedures.

### 1.1. IRT Likelihood Ratio Test Procedure

The IRT likelihood-ratio test procedure (Thissen, Steinberg, & Wainer 1993) is a parametric and model-based procedure for DIF detection. Both uniform and nonuniform DIF can be tested with this procedure. In the context of IRT, DIF can be defined in terms of item true score functions for both dichotomously and polytomously scored items (Kim et al. 2005). If there is no DIF, it is expected that the item true score function for the reference group and the one for the focal group will be the same. Otherwise, it is said that the item functions differentially between two groups.

In the IRT likelihood-ratio test procedure for DIF detection (Thissen, Steinberg, & Gerard 1986; Thissen, Steinberg, &Wainer 1988, 1993), the null hypothesis to be tested is that the item parameters between the reference group and the focal group do not differ. The difference in the item difficulty parameter between two groups is tested for the uniform DIF and the difference in the item discrimination parameter is tested for the nonuniform DIF. For the test of the null hypothesis of no DIF, two models are compared: a compact model and an augmented model. In the compact model, the item parameters for the common item or items across groups are constrained to be equal in the two groups. In the augmented model, the item parameters for the studied item are unconstrained and the remaining items are constrained to be equal in the two groups. Then the likelihood-ratio test statistic, $G^2$, is computed by the following equation,

$$G^2 = -2LL_C - (-2LL_A),$$

where $LL_C$ is the log likelihood for the compact model given the maximum likelihood estimates of the parameters of the compact model and $LL_A$ is the log likelihood for the augmented model given the maximum likelihood estimates of the parameters of the augmented model. The value of $G^2$ is distributed as the chi-square with the degrees of freedom equal to the difference in the number of parameters in the two models. If the result of the test is found to be significant, then it is said that the studied item exhibit DIF.

In one simulation study, the performance of IRT likelihood ratio test procedure in detecting DIF was investigated for 30 polytomously scored items with four ordered performance levels under partial credit model across the combination of three sample sizes and two ability matching conditions (Kim, Cohen, DiStefano &, Kim 1998). In another study, the performance of this procedure was investigated for 30 polytomously scored items with five ordered categories under graded response model across the same conditions (Kim & Cohen 1998). Both studies only examined the Type I error rates and not the power. As a result, they found that the Type I error rates were close to the nominal alpha levels across all simulation conditions. Ankenman, Witt, and Dunbar (1999) investigated the effects of sample size, ability matching, pattern of DIF, discrimination and threshold parameter values for the studied DIF item on Type I error rates and power of IRT likelihood ratio test procedure for a test with 20 dichotomously and 5 polytomously scored items under graded response model in a simulation study. They found similar results with the previous studies concerning the Type I error rates of IRT likelihood ratio test procedure. In addition, they found that the power of this procedure was affected by sample size, ability distribution, discriminating power of the studied item, and DIF pattern.

### 1.2. Ordinal Logistic Regression Procedure

Zumbo (1999) extended the logistic regression procedure for polytomously scored items. Zumbo uses cumulative logit model in his application of logistic regression procedure.

The logit for person *j* to score scoring category *k* or below is expressed as

$$\log it[P(Y_j \leq k | X_j, G_j)] = \log\left[\frac{P(Y_j \leq k)}{P(Y_j > k)}\right] = \beta_0 + \beta_1 X_j + \beta_2 G_j + \beta_3 (XG)_j ,$$

where $Y_j$ is the item response for person *j*; *k* is the response category; $X_j$ is the matching variable (i.e., total test score) of person *j*; $G_j$ is the dummy variable for group membership for person *j*, which is equal to 1 if the person belongs to group1 and 0 if the person belongs to group 2; ($XG_j$) is the interaction term between the observed ability level and the group membership for person *j*. $P(Y_j \leq k)$ is the probability of getting an item score less than or equal to *k* for person *j*; and $P(Y_j > k)$ is the probability of getting an item score greater than *k* for person *j*.Furthermore, $\beta_0$, $\beta_1$, $\beta_2$ and $\beta_3$ are the coefficients of the logistic regression DIF model. While $\beta_0$ is the intercept of the model, $\beta_1$, $\beta_2$, and $\beta_3$ are the slopes of the model. The item reflects uniform DIF if $\beta_2$ is nonzero and $\beta_3$ is zero whereas the item reflects nonuniform DIF if $\beta_3$ is nonzero.

The likelihood-ratio test statistic, $G^2$, is obtained to test the uniform and nonuniform DIF. $G^2$ for uniform test is computed by taking the difference of the values -2 times the log likelihood for the model with the matching variable (Model 1) and the values of -2 times the log likelihood for the model with the matching variable and group variable (Model 2). $G^2$ for nonuniform test is computed by taking the difference of the values -2 times the log likelihood for the model with the matching variable and group variable (Model 2) and the values of -2 times the log likelihood for the model with the matching variable, group variable, and interaction variable (Model 3). The value of $G^2$ is distributed as the chi-square with the degrees of freedom equal to the difference in the number of parameters in the two models.

Kristjansson, Aylesworth, McDowell, and Zumbo (2005) investigated the Type I error rates and power of ordinal logistic regression procedure in detecting DIF for 26 polytomously scored items with four score levels across the combination of two sample size ratios, two matching abilities, two skewness in ability distributions, and three studied item discriminations. They found that the Type I error rates of this procedure was in good control and the power to detect uniform and nonuniform DIF was above 0.90 across all simulation conditions.

## 2. METHOD

A series of Monte Carlo simulations with 200 replications was conducted for the IRT likelihood-ratio test and cumulative logit ordinal logistic regression DIF detection procedures under various simulation conditions. The IRT likelihood-ratio test procedure was implemented using the computer program IRTLRDIF (Thissen 2001) and the ordinal logistic regression procedure was implemented using the computer program SAS. Factors held constant and factors manipulated are described in detail in the following sections.

### 2.1. Factors Held Constant

Number of items (test length), number of scoring categories, number of DIF items, and ability distribution of the reference group (R) and the focal group (F) were held constant. Responses for 6 polytomously scored items with 4 scoring categories (0, 1, 2, and 3) were generated. Only one item was simulated as a DIF item in all simulation conditions. The ability parameters for both the reference and the focal group were sampled from a standard normal distribution *N*(0,1).

### 2.2. Factors Manipulated

**2.2.1. Sample size.** Three levels of sample size were investigated: small sample size (*N* = 600), medium sample size (*N* = 1200), and large sample size (*N* = 2400).

**2.2.2. Sample size ratio.** Two levels of sample size ratio were considered. Sample size ratio between the reference group and the focal group was set to 1:1 for the equal sample size conditions and 2:1 for the unequal sample size conditions. More specifically, it created conditions with *N* = 300R/300F, 400R/200F for the small sample size, *N* = 600R/600F, 800R/400F for the medium sample size, and *N* = 1200R/1200F, 1600R/800F for the large sample size.

**2.2.3. Magnitude of DIF.** Three levels of DIF magnitudes were considered: negligible DIF (0.32 logit), moderate DIF (0.43 logit), and large DIF (0.53 logit).

**2.2.4. DIF condition.** In this study, three DIF conditions were simulated by focusing on the differences in the between-category threshold parameters between the reference group and the focal group.

While the item discrimination parameter, *a*, was assumed to be the same for the reference and focal groups, the between-category threshold parameters, $b_k$, for the polytomous items varied across groups in the DIF conditions, If DIF is present in a polytomous item, one group will have a higher probability for the higher scoring categories and will score higher than the other group and one group will have a higher probability for the lower scoring probabilities and will score lower than the other group (French & Miller 1996).

Three DIF conditions were examined: low-shift DIF condition, high-shift DIF condition and balanced DIF condition (Chang, Mazzeo, & Roussos 1996). In this study, the condition where all the between-category threshold parameters have a constant amount of DIF against a particular group was not considered, since the constant-DIF condition is not common in actual testing situations.

In the low-shift DIF condition, the category threshold parameter between the lowest scoring category and the second scoring category were 0.32, 0.43, or 0.53 points higher for the focal group and the remaining between-category threshold parameters were the same for the two groups ($b_{i1F} = b_{i1R} + 0.32$, $b_{i2F} = b_{i2R}$, $b_{i3F} = b_{i3R}$; $b_{i1F} = b_{i1R} + 0.43$, $b_{i2F} = b_{i2R}$, $b_{i3F} = b_{i3R}$; $b_{i1F} = b_{i1R} + 0.53$, $b_{i2F} = b_{i2R}$, $b_{i3F} = b_{i3R}$).

In the high-shift DIF condition, the category threshold parameter between the third scoring category and the highest scoring category were 0.32, 0.43, or 0.53 points higher for the focal group and the remaining between-category threshold parameters were the same for both groups ($b_{i1F} = b_{i1R}$, $b_{i2F} = b_{i2R}$, $b_{i3F} = b_{i3R} + 0.32$; $b_{i1F} = b_{i1R}$, $b_{i2F} = b_{i2R}$, $b_{i3F} = b_{i3R} + 0.43$; $b_{i1F} = b_{i1R}$, $b_{i2F} = b_{i2R}$, $b_{i3F} = b_{i3R} + 0.53$).

In the balanced DIF condition, the first between-category threshold parameter was 0.32, 0.43, or 0.53 points higher for the focal group, the second between-category threshold parameter was the same for the two groups, and the third between-category threshold parameter was 0.32, 0.43, or 0.53 points higher for the reference group ($b_{i1F} = b_{i1R} + 0.32$, $b_{i2F} = b_{i2R}$, $b_{i3F} = b_{i3R} - 0.32$; $b_{i1F} = b_{i1R} + 0.43$, $b_{i2F} = b_{i2R}$, $b_{i3F} = b_{i3R} - 0.43$; $b_{i1F} = b_{i1R} + 0.53$, $b_{i2F} = b_{i2R}$, $b_{i3F} = b_{i3R} - 0.53$).

By completely corossing the four factors – 3 sample sizes, 2 sample size ratios, 3 DIF magnitudes, and 3 DIF conditions – 54 simulation conditions ($3 \times 2 \times 3 \times 3 = 54$ simulation conditions) were considered.

### 2.3. Data Generation Process

First, the ability parameter, $\theta$, was generated from the desired distribution for each examinee and was randomly assigned to each examinee. The probability of responding in category *k* or higher on item *i* for each examinee, $P^*_{ik}(\theta)$,was computed based on the Samejima'sgraded response model

(Samejima 1969; Samejima 1999) using the item parameters and the generated ability parameters. Then, a uniform random variate, *u*, from distribution (0, 1) was generated for each examinee, for each item. The generated random variate, *u*, was compared with the calculated probabilities. If the generated uniform random variate was less than the calculated probability at the category *k* but greater than the calculated probability at the category *k*+1, then the response was coded as *k*-1. For example, using the item parameters for the first item (item discrimination parameter of 1.500 and between-category threshold parameters of -1.750, -0.500, and 0.750) and the randomly generated ability parameter of 0.4418 for the first examinee, the item probabilities for the first item, for the first examinee is calculated as $P^*_{11}(0.4418) = 0.9640$, $P^*_{12}(0.4418) = 0.8042$, and $P^*_{13}(0.4418) = 0.3864$. If a uniform random variate of 0.6518 is generated for the first examinee, for the first item, a score of 1 is assigned to the first examinee for the first item since the generated uniform random variate is less than $P^*_{12}(0.4418) = 0.8042$ but greater than $P^*_{13}(0.4418) = 0.3864$.

### 2.4. Assessment of Type I Error Rate and Power

Type I error rate was investigated for items with no DIF. There were 5 non-DIF items in the polytomous test. Power was investigated for items with DIF. There was only 1 DIF item. The effects of the sample size, sample size ratio, magnitude of DIF, low-shift, high-shift, and balanced DIF conditions on the Type I error rates and power in detecting DIF items were examined. An alpha (α) level of 0.05 was considered as the nominal α level in the study. A widely used criterion of 0.80 was considered to evaluate the power results. Type I error rate less than or equal to 0.05 was considered to be good in control and Type I error rate greater than 0.05 was considered to be inflated. Power equal to or greater than 0.80 was considered to be high and power less than 0.80 was considered to be inadequate in detecting DIF.

## 3. RESULTS

### 3.1. Type I Error Study

In the Type I error study, the significance level of 0.05 was used as the nominal alpha level. The number of significant $G^2$s in the IRT likelihood-ratio test and ordinal logistic regression procedures for non-DIF items (5 non-DIF items for each simulation condition) over 200 replications was calculated at the nominal alpha level of 0.05.

**3.1.1. Low-shift DIF condition.** Type I error study results of the IRT likelihood-ratio test and the ordinal logistic regression procedures under low-shift DIF condition at the nominal alpha level of 0.05 are presented in Table 1.

**Table 1: Type I Error Rates under the Low-Shift DIF Condition**

| | | Likelihood-Ratio Test | | | Logistic Regression | | |
|---|---|---|---|---|---|---|---|
| | | DIF Magnitude | | | DIF Magnitude | | |
| Sample Size | Sample Size Ratio (R/F) | 0.32 | 0.43 | 0.53 | 0.32 | 0.43 | 0.53 |
| 600 | 300/300 | **0.044** | **0.048** | **0.053** | **0.045** | **0.049** | **0.053** |
| 600 | 400/200 | **0.053** | **0.052** | **0.049** | **0.047** | **0.052** | **0.057** |
| 1200 | 600/600 | **0.062** | **0.064** | **0.067** | **0.059** | **0.059** | **0.066** |
| 1200 | 800/400 | **0.040** | **0.084** | **0.070** | **0.054** | **0.076** | **0.081** |
| 2400 | 1200/1200 | **0.061** | **0.068** | **0.084** | **0.063** | **0.086** | **0.079** |
| 2400 | 1600/800 | **0.066** | **0.069** | **0.074** | **0.045** | **0.062** | **0.068** |

The Type I error rates ranged between 0.040 and 0.084 for the IRT likelihood-ratio test procedure, while they ranged between 0.045 and 0.086 for the ordinal logistic regression procedure at the nominal alpha level of 0.05. The empirical Type I error rates of both procedures were within or at their expected value of 0.05 when the sample size was small (*N* = 300R/300F or *N* = 400R/200F). They were slightly higher than nominal alpha level when the sample size was medium and the sample size ratio was equal (600/600). The empirical Type I error rates of both procedures deviated more for the combination of medium sample size with unequal sample size ratio (*N* = 800R/400F) and medium or high DIF magnitude (0.43 or 0.53) and for the combination of large sample size (*N* = 1200R/1200F or 1600R/800F) and medium to high DIF magnitude (0.43 or 0.53).

Under the equal sample size ratio condition (1:1), the Type I error rates of both procedures increased as the sample size increased. However, there was not any clear pattern for the unequal sample size ratio condition (2:1). The Type I error rates of both procedures increased as the DIF magnitude increased except when the sample size was small (*N* = 400R/200F) or medium (*N* = 800R/400F) for the IRT likelihood-ratio test procedure, when the sample size was large (*N* = 1200R/1200F) for the logistic regression procedure.

**3.1.2. High-shift DIF condition.** Type I error study results of the IRT likelihood-ratio test and the ordinal logistic regression procedures under high-shift DIF condition at the nominal alpha level of 0.05 are presented in Table 2.

**Table 2: Type I Error Rates under High-Shift DIF Condition**

| | | Likelihood-Ratio Test<br>DIF Magnitude | | | Logistic Regression<br>DIF Magnitude | | |
|---|---|---|---|---|---|---|---|
| Sample Size | Sample Size Ratio (R/F) | 0.32 | 0.43 | 0.53 | 0.32 | 0.43 | 0.53 |
| 600 | 300/300 | **0.044** | **0.046** | **0.043** | **0.048** | **0.048** | **0.059** |
| 600 | 400/200 | **0.048** | **0.048** | **0.055** | **0.051** | **0.048** | **0.049** |
| 1200 | 600/600 | **0.065** | **0.057** | **0.056** | **0.070** | **0.055** | **0.070** |
| 1200 | 800/400 | **0.048** | **0.054** | **0.054** | **0.051** | **0.056** | **0.056** |
| 2400 | 1200/1200 | **0.055** | **0.062** | **0.066** | **0.054** | **0.071** | **0.064** |
| 2400 | 1600/800 | **0.059** | **0.071** | **0.067** | **0.038** | **0.063** | **0.062** |

The Type I error rates of the IRT likelihood-ratio test procedure ranged between 0.043 and 0.071, and the Type I error rates of the ordinal logistic regression procedure ranged between 0.038 and 0.071 at the nominal alpha level of 0.05. In general, the Type I error rates of both procedures were close to the expected alpha level of 0.05. The Type I error rates were slightly higher than their expected value when the sample size was medium (*N* = 600R/600F) or large (*N* = 1200R/1200F or *N* = 1600R/800F). When compared with the low-shift DIF condition, the Type I error rates were lower for the high-shift DIF condition at most conditions. There were not any clear patterns of the Type I error rates across sample sizes or DIF magnitudes for both procedures.

**3.1.3. Balanced DIF condition.** Type I error study results of the IRT likelihood-ratio test and the ordinal logistic regression procedures under balanced DIF condition at the nominal alpha level of 0.05 were presented in Table 3.

**Table 3: Type I Error Rates under Balanced DIF Condition**

| | | Likelihood-Ratio Test<br>DIF Magnitude | | | Logistic Regression<br>DIF Magnitude | | |
|---|---|---|---|---|---|---|---|
| Sample Size | Sample Size Ratio (R/F) | 0.32 | 0.43 | 0.53 | 0.32 | 0.43 | 0.53 |
| 600 | 300/300 | **0.047** | **0.049** | **0.072** | **0.061** | **0.059** | **0.066** |
| 600 | 400/200 | **0.070** | **0.057** | **0.068** | **0.068** | **0.045** | **0.058** |
| 1200 | 600/600 | **0.048** | **0.061** | **0.048** | **0.053** | **0.062** | **0.051** |
| 1200 | 800/400 | **0.046** | **0.050** | **0.068** | **0.053** | **0.051** | **0.044** |
| 2400 | 1200/1200 | **0.056** | **0.063** | **0.077** | **0.042** | **0.051** | **0.057** |
| 2400 | 1600/800 | **0.069** | **0.066** | **0.075** | **0.049** | **0.053** | **0.052** |

The Type I error rates of the IRT likelihood-ratio test procedure ranged between 0.046 and 0.077, and the Type I error rates of the ordinal logistic regression procedure ranged between 0.042 and 0.068 at the nominal alpha level of 0.05. In general, the Type I error rates of both procedures were close to their expected value of 0.05. The logistic regression procedure provided better control of the Type I error rates than the IRT likelihood-ratio test procedure with the combination of large sample size (*N* = 1200R/1200F or *N* = 1600R/800F) and large DIF magnitude (0.53). There were not any clear patterns in the Type I error rates across sample sizes and DIF magnitudes for both procedures.

### 3.2. Power Study

For power study, the number of significant $G^2$s in both procedures for DIF items (Only 1 DIF item for each simulation condition) over 200 replications was calculated at the nominal alpha level of 0.05.

**3.2.1. Low-shift DIF condition.** Power study results of the IRT likelihood-ratio test and the ordinal logistic regression procedures under low-shift DIF condition at the nominal alpha level of 0.05 are presented in Table 4.

**Table 4: Power under Low-Shift DIF Condition**

| | | Likelihood-Ratio Test | | | Logistic Regression | | |
|---|---|---|---|---|---|---|---|
| | | DIF Magnitude | | | DIF Magnitude | | |
| Sample Size | Sample Size Ratio (R/F) | 0.32 | 0.43 | 0.53 | 0.32 | 0.43 | 0.53 |
| 600 | 300/300 | **0.470** | **0.770** | **0.900** | **0.120** | **0.200** | **0.310** |
| 600 | 400/200 | **0.385** | **0.650** | **0.860** | **0.130** | **0.185** | **0.275** |
| 1200 | 600/600 | **0.790** | **0.985** | **1.000** | **0.260** | **0.455** | **0.585** |
| 1200 | 800/400 | **0.680** | **0.935** | **1.000** | **0.170** | **0.355** | **0.495** |
| 2400 | 1200/1200 | **0.995** | **1.000** | **1.000** | **0.390** | **0.640** | **0.870** |
| 2400 | 1600/800 | **0.945** | **1.000** | **1.000** | **0.380** | **0.660** | **0.780** |

The power of the IRT likelihood-ratio test procedure ranged between 0.385 and 1.000, and the power of the ordinal logistic regression procedure ranged between 0.120 and 0.870 at the nominal alpha level of 0.05. Thus, the IRT likelihood-ratio test procedure was generally more powerful than the ordinal logistic regression procedure at each condition.

Under the equal sample size ratio (1:1) or unequal sample size ratio (2:1) conditions, the power of IRT likelihood-ratio test and ordinal logistic regression procedures increased as the sample size increased. The power of these procedures increased as the DIF magnitude increased. The power of the ordinal logistic regression procedure exceeded 0.80 only for the combination of large sample size with equal sample size ratio (*N* = 1200R/1200F) and high DIF magnitude (0.53). The power of the ordinal logistic regression procedure did not exceed the 0.8 criterion for most of the other conditions. On the other hand, IRT likelihood-ratio test procedure was demonstrated to be powerful in most of the conditions. It had perfect power when the sample size was medium (*N* = 600R/600F or *N* = 800R/400F) or high (*N* = 1200R/1200F or 1600/800) and the DIF magnitude was large (0.53) or when the sample size was large (*N* = 1200R/1200F or *N* = 1600R/800F) and the DIF magnitude was medium (0.43). Its power was very low when the sample size was small (*N* = 300R/300F or *N* = 400R/200F) and the DIF magnitude was small (0.32).

**High-shift DIF condition.** Power study results of the IRT likelihood-ratio test and the ordinal logistic regression procedures under high-shift DIF condition at the nominal alpha level of 0.05 are presented in Table 5.

**Table 5: Power under High-Shift DIF Condition**

| | | Likelihood-Ratio Test<br>DIF Magnitude | | | Logistic Regression<br>DIF Magnitude | | |
|---|---|---|---|---|---|---|---|
| Sample Size | Sample Size Ratio (R/F) | 0.32 | 0.43 | 0.53 | 0.32 | 0.43 | 0.53 |
| 600 | 300/300 | **0.285** | **0.510** | **0.740** | **0.115** | **0.180** | **0.235** |
| 600 | 400/200 | **0.295** | **0.500** | **0.695** | **0.095** | **0.160** | **0.185** |
| 1200 | 600/600 | **0.635** | **0.905** | **0.965** | **0.180** | **0.300** | **0.430** |
| 1200 | 800/400 | **0.600** | **0.880** | **0.895** | **0.165** | **0.245** | **0.350** |
| 2400 | 1200/1200 | **0.915** | **0.995** | **1.000** | **0.310** | **0.560** | **0.740** |
| 2400 | 1600/800 | **0.870** | **0.995** | **1.000** | **0.300** | **0.425** | **0.575** |

The power of the IRT likelihood-ratio test procedure ranged between 0.285 and 1.000, and the power of the ordinal logistic regression procedure ranged between 0.095 and 0.740 at the nominal alpha level of 0.05. The IRT likelihood-ratio test procedure was more powerful in detecting DIF than the logistic regression procedure in all conditions. The power of the logistic regression procedure did not exceed 0.80 in any condition, and it was interpreted inadequate in detecting DIF at all conditions.

Under the equal sample size ratio (1:1) or unequal sample size ratio (2:1) conditions, the power of the IRT likelihood-ratio test and the ordinal logistic regression procedures increased as the sample size or the DIF magnitude increased. The IRT likelihood-ratio test procedure was very powerful when the sample size was large (*N* = 1200R/1200F or *N* = 1600R/800F). It had perfect power with the combination of the large sample size (*N* = 1200R/1200F or *N* = 1600R/800F) and the large magnitude of DIF (0.53). On the other hand, the power of the IRT likelihood-ratio test procedure was below 0.80 for the small sample size (*N* = 300R/300F or *N* = 400R/200F) or for the combination of medium sample size (*N* = 600R/600F or *N* = 800R/400F) and small DIF magnitude (0.32) conditions.

**Balanced DIF condition.** Power study results of the IRT likelihood-ratio test and logistic regression procedures for polytomous test items under balanced DIF condition at the nominal alpha level of 0.05 are reported in Table 6.

**Table 6: Power under Balanced DIF Condition**

| | | Likelihood-Ratio Test<br>DIF Magnitude | | | Logistic Regression<br>DIF Magnitude | | |
|---|---|---|---|---|---|---|---|
| Sample Size | Sample Size Ratio (R/F) | 0.32 | 0.43 | 0.53 | 0.32 | 0.43 | 0.53 |
| 600 | 300/300 | **0.805** | **0.950** | **1.000** | **0.080** | **0.070** | **0.050** |
| 600 | 400/200 | **0.730** | **0.950** | **1.000** | **0.060** | **0.040** | **0.070** |
| 1200 | 600/600 | **0.990** | **1.000** | **1.000** | **0.055** | **0.055** | **0.060** |
| 1200 | 800/400 | **0.960** | **1.000** | **1.000** | **0.055** | **0.065** | **0.085** |
| 2400 | 1200/1200 | **1.000** | **1.000** | **1.000** | **0.040** | **0.055** | **0.045** |
| 2400 | 1600/800 | **1.000** | **1.000** | **1.000** | **0.065** | **0.085** | **0.080** |

The power of the IRT likelihood-ratio test procedure ranged between 0.730 and 1.000, and the power of the ordinal logistic regression procedure ranged between 0.040 and 0.085 at the nominal alpha level of .05. The power of the ordinal logistic regression procedure was extremely low in all conditions. On the other hand, the IRT likelihood-ratio test procedure was very powerful in almost all conditions. It provided perfect power when the sample size was large (*N* = 1200R/1200F or *N* = 1600R/800F) and the magnitude of DIF was small (0.32), when the sample size was medium (*N* = 600R/600F or *N* = 800R/400F) or large (*N* = 1200R/1200F or *N* = 1600R/800F) and the magnitude of DIF was medium (0.43), or when the magnitude of DIF was large (0.53).

The power of the IRT likelihood-ratio test procedure was higher for the balanced DIF condition than for the low-shift and high-shift DIF conditions. On the other hand, the power of the ordinal logistic regression procedure was very low for the balanced DIF condition.

## 4. DISCUSSION and CONCLUSION

Type I error rates of the IRT likelihood-ratio test and the ordinal logistic regression procedures below or at the nominal alpha level of 0.05 was considered as to be good in control and the Type I error rates above 0.05 was considered to be inflated. Power at or above 0.80 was considered to be high, and power below 0.80 was considered as to be inadequate.

When the low-shift DIF condition was considered, the Type I error rates of the IRT likelihood-ratio test and the ordinal logistic regression procedures were within or at the nominal alpha level for small sample size. Type I error rates were higher than nominal alpha level for medium or large sample sizes, especially, when the sample size was medium or large, for either equal or unequal sample size ratio conditions, and the magnitude of DIF was medium or large. When compared with the low-shift DIF condition, the high-shift DIF condition provided lower Type I error rates for both procedures at most conditions. As opposed to the low-shift and high-shift DIF conditions, the Type I error rates were higher for small sample size under the balanced DIF condition. Type I error rates of the IRT likelihood-ratio test procedure was also high for large sample size. On the other hand, Type I error rates of the ordinal logistic regression procedure were in good control for this sample size.

The power of the IRT likelihood-ratio test procedure was very high for the low-shift DIF condition, except for the combination of small or medium sample size and small or medium DIF magnitude conditions. In addition to these exceptional conditions, the combination of small sample size and large DIF magnitude conditions had low power in high-shift DIF condition. The power of high-shift DIF condition was generally lower than the power of low-shift DIF condition across all conditions. For balanced DIF condition, the IRT likelihood-ratio test procedure was very powerful at all conditions even when the sample size and the magnitude of DIF were small except for the unequal sample size ratio condition. Overall, the balanced DIF condition provided best detection rates for the IRT likelihood-ratio test procedure. It was followed by the low-shift DIF condition.

The power of the ordinal logistic regression procedure was unacceptably low at most conditions for all 3 DIF conditions. The power of the ordinal logistic regression procedure exceeded 0.80 only for the combination of large sample size with equal sample size ratio and large DIF magnitude condition under low-shift DIF condition. The power of this procedure was extremely low under balanced DIF condition. The power did not exceed 0.10 at all conditions. These findings are not consistent with the findings of Kristjansson, Aylesworth, McDowell, and Zumbo (2005). In their study, the power was above 0.90 for all simulation conditions. The low power values in this study may be because of the number of items.

In this study, only uniform DIF was examined with the IRT likelihood-ratio test and the ordinal logistic regression procedures for the simulation study. It is also possible to examine nonuniform DIF with these two DIF detection procedures.

The Type I error rates and power of the IRT likelihood-ratio test and the ordinal logistic regression procedures were evaluated based on the known parameters with the simulation study. The generalization of the results is limited to the conditions that were used in the study.

In the simulation study, the sample size ratio between the reference group and the focal group was set to 1:1 and 2:1. However, there might be the situations that the sample size ratio between these two groups is more distinct in actual testing.

Number of items (test length), number of scoring categories, and number of DIF items were held constant in the simulation study. 6 polytomously scored items with 4 scoring categories were generated. However, large scale assessment tests might have more or less polytomously scored items. Test length might have an important effect on the item parameter estimation and on the DIF detection. Longer tests provide more information about examinees and this might increase the accuracy of the parameter estimation and the power of the DIF detection. Polytomous items in large scale assessment test might also have different number of scoring categories. For example, polytomous items with 3 scoring categories for short-response, 5 scoring categories for extended response items. The effect of the test length and number of scoring categories for polytomous items might be investigated. Only one DIF item was generated for each test in the simulation study. It might be interesting to examine the effect of the number of DIF items in a test.

## 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*(1), 67-91.

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodnes-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement*, *36*(4), 277-300.

Chang, H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement*, *33*(3), 333-353.

Cohen, A. S., Kim, S., & Baker, F. B.(1993).Detection of differential item functioning in the graded response model. *Applied Psychological Measurement*, *17*(4), 335-350.

Dorans, N. J. & Kulick E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement*, *23*(4), 355-368.

Dorans, N. J., & Schmitt, A. P. (1993). *Constructed response and differential item functioning: A pragmatic approach*. (ETS-RR-91-47). Princeton, NJ: Educational Testing Service.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement*, 33(3), 315-332.

Holland, P.W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds), *Test validity* (pp.129-145). Hillsdale, NL: Erlbaum.

Kim, S., Cohen, A. S., DiStefano, C. A., & Kim, S. (1998). *An investigation of the likelihood ratio test for detection of differential item functioningunder the partial credit models*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Kim, S. & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement*, *22*(4), 345-355.

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement*, *65* (6), 935-953.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719-748.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement*, 30(2), 107-122.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.

Samejima, F. (1999). *General graded response model*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Shealy, R. T., & Stout, W. F. (1993). An item response theory model for test bias and differential test functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp.197-239). Hillsdale, NJ: Erlbaum.

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

Thissen, D. (2001). IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning. University of North Carolina at Chapel Hill: L. L. Thurstone Psychometric Laboratory.

Thissen, D., Steinberg, L., & Gerard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, *99*(1), 118-128.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun(Eds.), *Test validity* (pp.147-169). Hillsdale, NJ: Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp.67-113). Hillsdale, NJ: Erlbaum.

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, Ontario, Canada: Directorate of Human resources research and Evaluation, Department of National Defence.

Zwick, R., Donoghue, J. R., & Grima, A. (1993a). *Assessing differential item functioning in performance tests* (ETS-RR-93-14). Princeton, NJ: Educational Testing Service.

Zwick, R., Donoghue, J. R., & Grima, A. (1993b). Assessing differential item functioning for performance tests. *Journal of Educational Measurement*, *30*(3), 233-251.

## Genişletilmiş Özet

Eğitim alanında sadece çok kategorili puanlanan maddelerden oluşan testleri ya da hem iki kategorili puanlanan, hem de çok kategorili puanlanan maddelerden oluşan testleri görmek mümkündür. Bu nedenle, değişen madde fonksiyonu (DMF) gösteren çok kategorili puanlanan maddelerin belirlenmesi, DMF gösteren iki kategorili puanlanan maddelerin belirlenmesi kadar önemlidir. Madde tepki kuramı (MTK) çerçevesinde ele alındığında, çok kategorili puanlanan bir madde için değişen madde fonksiyonu, farklı gruplarda aynı yetenek düzeyine sahip bireylerin kestirilen puan fonksiyonlarının birbirine eşit olmaması şeklinde tanımlanır (Cohen, Kim, & Baker 1993; Kim & Cohen 1998). DMF gösteren maddeler, test puanlarının geçerli bir şekilde yorumlanmasına ve kullanılmasına yönelik bir tehdit oluşturmaktadırlar. Bu nedenle, testte DMF gösteren maddelerin tespit edilmesi önemlidir.

İki kategorili puanlanan maddelerden oluşan testlerde DMF gösteren maddeleri belirlemede kullanılan çok sayıda istatistiksel yöntem mevcuttur. Bunlardan yaygın olarak kullanılanları, Mantel-Haenszel (MH: Holland & Thayer 1988), standardizasyon (Dorans & Kulick 1986), lojistik regresyon (LR: Swaminathan & Rogers 1990), simültane madde yanlılığı testi (SIBTEST: Shealy & Stout 1993) yöntemleri ile madde tepki kuramına bağlı yöntemlerdir (Thissen, Steinberg, & Gerard 1986; Thissen, Steinberg, & Wainer 1988, 1993). Bu yöntemlerden birçoğu çok kategorili puanlanan maddelerden oluşan testlerde DMF gösteren maddelerin belirlenmesinde kullanılmak üzere genişletilmişlerdir.

Bu çalışmada, MTK olabilirlik oranı testi ve kümülatif lojit ordinal lojistik regresyon yöntemlerinin çok kategorili puanlanan maddelerin değişen madde fonksiyonunu belirlemedeki performansları karşılaştırılmıştır. Bu iki yöntemin çok kategorili puanlanan maddelerin değişen madde fonksiyonunu belirlemedeki etkinlikleri hakkında literatürde yeterli araştırma yoktur.

Bu çalışmada kullanılan veriler, Monte Carlo simülasyon yöntemi ile üretilmiştir. Simülasyon yöntemi ile çalışmadaki birey sayısı, grupların yetenek dağılımları, madde sayısı, madde parametreleri, DMF gösteren maddeler ve bu maddelerin gösterdiği DMF büyüklükleri gibi faktörler önceden belirlenerek madde puanları yapay olarak oluşturulur. Bu çalışmada ele alınan simülasyon koşulları, örneklem büyüklüğü (küçük (600 birey), orta (1200 birey) ve büyük (2400 birey)), referans grup (R) ile odak grup (O) arasındaki örneklem büyüklüğü oranı (1:1 (300R/300O; 600R/600O; 1200R/1200O) ve 2:1 (400R/200O; 800R/400O; 1600R/800O)), DMF büyüklüğü (küçük (0.32 lojit), orta (0.43 lojit) ve büyük (0.53 lojit)) ve DMF durumu (alta kayan (birinci ve ikinci puan kategorileri arasındaki kategori eşik parametresinin odak grup için daha büyük olduğu durum), üste kayan (üçüncü ve dördüncü puan kategorileri arasındaki kategori eşik parametresinin odak grup için daha büyük olduğu durum)ve dengeli (birinci kategoriler arası eşik parametresinin odak grup için daha büyük, üçüncü kategoriler arası eşik parametresinin ise referans grup için daha büyük olduğu durum)) olmak üzere dört faktör altında toplanmıştır. Böylece 54 simülasyon koşulu (3 örneklem büyüklüğü, 2 örneklem büyüklüğü oranı, 3 DMF büyüklüğü ve 3 DMF durumu) üretilmiştir ve herbir simülasyon koşulu için MTK olabilirlik oranı testi analizleri ve ordinal lojistik regresyon analizleri 200'er kere tekrar edilmiştir. MTK olabilirlik oranı testi analizleri IRTLRDIF yazılımı kullanılarak (Thissen 2001), ordinal lojistik regresyon analizleri ise SAS yazılımı kullanılarak gerçekleştirilmiştir.

Bu iki yöntemin performansı simülasyon ile oluşturulan farklı koşullar altında tip I hata oranları ve güçleri ele alınarak karşılaştırılmıştır. Tip I hata çalışması için 0.05 değeri nominal alfa düzeyi olarak kullanılmıştır. Böylece MTK olabilirlik oranı ve ordinal lojistik regresyon yöntemlerinin tip I hata oranlarının 0.05 değerinden küçük veya 0.05 değerine eşit olması, hata oranlarının kontrol altında olduğu, 0.05 değerinin üzerinde olması ise, hata oranlarının yüksek olduğu şeklinde yorumlanmıştır. Güç çalışması için 0.80 değeri kriter olarak kullanılmıştır. Böylece her iki yöntemin gücünün 0.80 değerine eşit olması veya 0.80 değerinden büyük olması gücün yüksek olduğu, 0.80 değerinin altında olması ise gücün yetersiz olduğu şeklinde yorumlanmıştır.

MTK olabilirlik oranı testi ve ordinal lojistik regresyon yöntemlerinin tip I hata oranları genel olarak bütün simülasyon koşulları altında iyi kontrol sağlamıştır. MTK olabilirlik oranı testinin gücü orta veya büyük örneklem büyüklüğü ve orta veya büyük DMF büyüklüğü için 0.80'in üzerinde bulunmuştur. Bu yöntemin gücü örneklem büyüklüğü veya DMF büyüklüğü arttıkça artmıştır. Bu yöntem özellikle dengeli değişen madde fonksiyonu durumu için bütün koşullar altında yüksek güç değerleri sağlamıştır. Diğer yandan, ordinal lojistik regresyon yönteminin gücü bütün simulasyon koşulları için özellikle de dengeli değişen madde fonksiyonu durumu için kabul edilemez derecede düşük çıkmıştır. Bu yöntem sadece büyük örneklem büyüklüğü ve büyük DMF koşulu altında yüksek bir güç değeri sağlamıştır.

Bulguların genellenebilirliği bu çalışmada ele alınan koşullarla sınırlıdır. Testteki madde sayısı, puanlama kategorilerinin sayısı ve DMF maddelerinin sayısı simülasyon çalışmasında sabit tutulmuştur. Sadece 1 tane DMF maddesi içeren 6 tane 4 kategorili puanlanan madde üretilmiştir. Halbuki, geniş ölçekli testlerde bu çalışmada kullanılan madde sayısından daha az veya daha fazla madde olabilir. Test uzunluğunun DMF belirlemede önemli bir etkisi olabilir. Uzun testler bireyler hakkında daha fazla bilgi sağladıkları için bu durum parametre kestiriminin doğruluğunu ve DMF belirlemenin gücünü artırabilir. Ayrıca büyük ölçekli testlerde çok kategorili puanlanan maddelerin kategori sayısı bu çalışmada kullanılan kategori sayısından farklı olabilir. Örneğin, kısa cevap gerektiren maddeler için 3 kategorili puanlamaya, uzun cevap gerektiren maddeler için ise 5 kategorili puanlamaya rastlanabilir. Test uzunluğunun ve çok kategorili maddelerin kategori sayısının DMF belirlemedeki etkisi incelenebilir.