Educational Sciences: Theory & Practice • 14(6) • 2186-2193

www.edam.com.tr/estp
DOI: 10.12738/estp.2014.6.2165

Comparing Performances (Type I error and Power) of IRT Likelihood Ratio SIBTEST and Mantel–Haenszel Methods in the Determination of Differential Item Functioning

Kübra ATALAY KABASAKAL[a]
Hacettepe University

Nihan ARSAN[b]

Bilge GÖK[c]
Hacettepe University

Hülya KELECİOĞLU[d]
Hacettepe University

Abstract

This simulation study compared the performances (Type I error and power) of Mantel-Haenszel (MH), SIBTEST, and item response theory-likelihood ratio (IRT-LR) methods under certain conditions. Manipulated factors were sample size, ability differences between groups, test length, the percentage of differential item functioning (DIF), and underlying model used to generate data. Results suggest that SIBTEST had the highest Type I error in the detection of uniform DIF, but MH had the highest power under all conditions. In addition, the percentage of DIF and the underlying model appear to have influenced the Type I error rate of IRT-LR. Ability differences between groups, test length, the percentage of DIF, model, and the interactions between ability differences*percentage of DIF, ability differences*test length, test length*percentage of DIF, test length*model affected the SIBTEST methods' Type I error rate. In the MH procedure, effective factors for Type I error rate were: sample size, test length, the percentage of DIF, ability differences*percentage of DIF, ability differences*model, and ability differences*percentage of DIF*model. No factors were effective on the power of SIBTEST and MH, but the underlying model had a significant effect on the IRT-LR power rate.

Keywords

Differential Item Functioning, SIBTEST, Item Response Theory-Likelihood Ratio, Mantel-Haenszel, Type I Error, Power.

a **Kübra ATALAY KABASAKAL, Ph.D.,** is currently a doctor of Measurement and Evaluation in Education. Her research interests include differential item functioning, test equating and multilevel modeling. *Correspondence:* Hacettepe University, Faculty of Education, Department of Educational Sciences, Ankara, Turkey. Email: katalay@hacettepe.edu.tr

b Nihan ARSAN, Ph.D., is currently a doctor of Measurement and Evaluation in Education. Email: nihanarsan@gmail.com

c Bilge GÖK, Ph.D., is currently a doctor of Measurement and Evaluation in Education. Contact: Hacettepe University, Faculty of Education, Department of Elementary Education, Ankara, Turkey. Email: bilgeb@hacettepe.edu.tr

d Hülya KELECİOĞLU, Ph.D., is currently a professor of Measurement and Evaluation in Education. Contact: Hacettepe University, Faculty of Education, Department of Educational Sciences, Ankara, Turkey. Email: hulyaebb@hacettepe.edu.tr

Differential item functioning (DIF) is an essential step in gathering score validity evidence. It exists when examinees with the same ability level have different probabilities of success on a given item (Holland & Wainer, 1993). DIF can be evidence of item bias, and biased items decrease a test's validity and cause unfair scoring. In an educational context, results of DIF studies can be used to organize more valid, fairer measurements. There are too many statistical procedures to detect DIF (Narayanan & Swaminathan, 1994), but there are several techniques designed to determine whether a test item is functioning differentially. Some of these techniques are based on classical test theory (CTT) and others on item response theory (IRT). Mantel-Haenszel (MH) (Holland & Thayer, 1988), logistic regression (LR) (Swaminathan & Rogers, 1990), and SIBTEST (Shealy & Stout, 1993) are based on CTT. Examples for IRT based methods are Lord's chi-square test, Raju's area measures, and likelihood ratio. DIF is examined by comparing item response distribution for two different groups of examinees with equal ability levels. Examinees with the same knowledge must respond similarly to test questions, regardless of their group membership. Differences in distributions are interpreted as DIF (Steinberg & Thissen, 2006).

CTT based methods compare groups' score distributions, but in IRT methods, probabilities of responding correctly to the items are compared. IRT methods are based on models, and the comparison parameters are changed according to the models. For example: in 1PLM, groups are compared with respect to b- item difficulty parameter; in 2PLM, a-item discrimination and b parameters are used for comparison. Between groups, b parameter differences indicate uniform DIF; differences in a parameter indicate non-uniform DIF.

Performances of DIF detection methods are not the same. IRT based methods are theoretically powerful, but large samples are required. Practically, satisfying this condition is difficult (Narayanan & Swaminathan, 1994). DIF studies regarding methods' performances have found that several factors, e.g., test length, sample size, test group size, group mean difference, standard deviation difference, distribution of difference, and interaction of these factors can be affected (Ackerman & Evans, 1992; Finch, 2005; Finch & French, 2007; Kim, 2010; Narayanan & Swaminathan, 1994; Prieto, Barbero, & San Luis, 1997; Rogers & Swaminathan, 1993; Roussos & Stout, 1996; Shealy & Stout, 1993).

In this research, Type I error rate and power of the MH procedure, IRT-LR, and SIBTEST methods are investigated based on sample sizes, ability differences between groups, test length, percentage of DIF, and the underlying model (2PL and 3PL). Below, these three methods are explained in detail.

DIF Detection Methods

Studies on DIF in Turkey have mostly used MH and LR methods (Bakan-Kalaycıoğlu ve Kelecioğlu, 2011; Bekçi, 2007; Çepni, 2011; Karakaya, 2012; Karakaya & Kutlu, 2012). In the present study, MH, SIBTEST, and IRT-LR methods were used. The LR method was not used for two main reasons: (1) in some studies, the LR method gave the same results as the MH method (Ankenmann, Witt, & Dunbar, 1996; DeMars, 2009; Vaughn & Wang, 2010), and (2) the error rate of the LR method was very high, and its statistical power, lower (Dainis, 2008; Hidalgo & Lopez-Pina, 2004; Jodoin & Gierl, 2001; Li, Brooks, & Johanson, 2012). In addition to this, the main weakness of the LR method in DIF determination is a tendency to produce higher Type I error (Li & Stout, 1996; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990).

Mantel-Haenszel (MH): The MH procedure is a common method in DIF detection. This procedure was developed to detect uniform DIF, and it is based on chi-square statistics. Also, the MH procedure is based on estimating the probability of a member of the reference or focal group (Agresti, 1984).

The Mantel-Haenszel procedure follows a three-dimension probability table that tests the independence of two variables (Vaughn & Wang, 2010). Holland and Thayer adapted the MH procedure to detect DIF (Dorans & Holland, 1993). The MH procedure is suitable only for uniform DIF, and probability ratio is calculated first. Probability ratio (α_{MH}) can change between 0 and ∞. Logarithmic transformation of probability ratio is calculated for interpretive purposes, and ΔMH value is obtained. The DIF level of the item is interpreted based on ΔMH:

- Type A items—negligible DIF: items with $|\Delta MH_i| < 1$
- Type B items—moderate DIF: items with $1 \leq |\Delta MH_i| < 1{,}5$
- Type C items—large DIF: items with $|\Delta MH_i| \geq 1{,}5$ (Zieky, 1993).

Item Response Theory Likelihood Ratio (IRTLR): This is a method that mainly compares the likelihood ratios of the models based on IRT (Thissen, Steinberg, & Wainer, 1993). DIF detection in likelihood ratio is tested with null hypothesis based on the comparison of item parameters of focal and reference groups. Compact and augmented groups are formed to test this null hypothesis, and the likelihood ratios of these models are compared. In a compact model, item parameters are considered equal across reference and focal groups. In an augmented model, item parameters of the item i differed while they were equal for other items in reference and focal groups. Then, logarithmic transformations were applied to both models, and G^2 value was calculated (Thissen, 2001). This process was replicated for all items.

$G^2 = -2LL_c - (-2LL_a)$

L_c: likelihood ratio of compact model

L_a: liklihood ratio of augmented model

If the value of G^2, which shows the distribution of X^2, is significant, sequential tests are applied to test the differences of item parameters. Each test of item parameters was conducted one by one.

SIBTEST: SIBTEST is a non-parametric latent variable model (Cheng, 2005; Potenza & Dorans, 1995), designed to detect uniform DIF. Li and Stout (1993) developed crossed SIBTEST to detect non-uniform DIF. To detect differences between groups, the SIBTEST method compares the correct response probabilities of the reference and focal groups for an item i (Camilli, 2006).

Each of the DIF detection methods has advantages and disadvantages under different circumstances. Recently, latent response models have become more popular because they are used widely in IRT scaling and in large-scale tests. The detection of DIF with appropriate methods has an important role in preparing unbiased, reliable tests. Thus, for accurate decisions, avoiding items with DIF is very important. For these reasons, determining which method or methods are more accurate under certain circumstances provides the opportunity to obtain accurate decisions. MH is widely used in the research literature. In addition, IRT-LR and SIBTEST are becoming researchers' focus of interest because of their invariance properties. Besides that, these methods provide an analytic framework for determining real group differences without error and bias (Shepard, Camilli, & Williams, 1984).

Research results showed that the Type I errors and powers of the methods are affected by too many conditions, and the methods give different results under different conditions (Ackerman & Evans, 1992; Finch, 2005; Narayanan & Swaminathan, 1994; Prieto et al., 1997; Rogers & Swaminathan, 1993; Roussos, 1992; Roussos & Stout, 1996; Shealy & Stout 1993; Uttaro & Millsap, 1994; Zwick, 1990).

Atalay, Gök, Kelecioğlu, and Arsan (2012) compared the MH, LR, IRT-LR, and SIBTEST methods to determine items with DIF in a simulation study, using the following simulation conditions: sample sizes (equal sample sizes for focal and reference groups 400–400; 1500–1500), trait distributions [(N(0,1) and N(0,1); (N(0,1) and N(0.5,1)], and items with DIF (5% and 10%). Study results found IRT-LR more sensitive than SIBTEST, and SIBTEST more sensitive than LR and MH. Moreover, MH was more sensitive than the LR method for determining DIF items. Methods were also compared to determine uniform and non-uniform DIF items, and these four methods consistently determined uniform DIF. Additionally, LR, SIBTEST, and IRT-LR consistently determined non-uniform DIF.

Erdem Keklik (2012) also compared the MH, IRT-LR, and LR methods to determine uniform DIF in a simulation study involving trait distributions and sample sizes. The results indicated that IRT-LR was better than MH and LR methods for controlling Type I error in differentiated trait distributions. But when the trait distributions were normally distributed, the MH and LR methods were similar and had lower Type I error than IRT-LR.

The present study's main aim was to determine the power of the four approaches for detection of DIF under a variety of conditions. In this framework, DIF detection methods such as Mantel-Haenszel, IRT-LR, and SIBTEST, were compared under the following different conditions: sample sizes, trait distributions of groups, length of tests, ratio of items with DIF, and underlying model. Type I and power of the DIF methods were compared under these conditions.

Method

As a simulation study, this research compared DIF detection methods under different conditions with respect to their power and Type I error. Previous research found that DIF methods were affected by various variables, such as sample sizes, trait distributions of comparison groups, focus and

reference groups, length of tests, the proportion of items with DIF, and differential test functioning magnitudes (Clauser, Mazor, & Hambleton, 1993; Finch & French, 2007; Narayanan & Swaminathan, 1996). The present study's conditions were selected according to their impact on DIF detection methods. Beyond that, characteristics of real data applications/conditions in Turkey were considered. Thus, the present study's results will provide proposals for the determination of analytical methods of examinations in Turkey. The simulation conditions are presented below.

Simulation Conditions

Simulation was conducted to examine the power and Type I error rates of four DIF detection methods with independent variables manipulated as follows: sample sizes, differences in trait distributions, test length, items with DIF, and model type. Impact sizes and types of DIF are factors that affect DIF detection methods. Nevertheless, these factors remain constant in this study, and the reference group is favored in all conditions.

Sample Sizes: Sample size is one important factor for detection of DIF. In nonparametric methods, the power of non-equal sample sizes in detection of DIF is higher than the presence of equal sample sizes (Kristjansson, Aylesworth, McDowell, & Zumbo, 2005). Simulation studies showed that, to reach better parameter estimates, the minimum sample sizes for MH and LR methods are 200–250 per group (Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993). However, IRT methods require large sample sizes (i.e., 1000 or more) for accurate estimations (Shepard, Camilli, & Averill, 1981). In the present study, sample size was simulated at 2400. Two conditions were created: equal focal and reference groups (1200:1200) and unequal focal and reference groups (800:1600).

Trait Distributions: In this study, the second factor manipulated was the difference in trait distribution between focal and reference groups. In a real data setting, it is difficult to find groups with the same trait distribution, and trait differences can influence DIF detection (Jodoin & Gierl, 2001; Narayanan & Swaminathan, 1994). Thus, two conditions were simulated. In the first condition, the reference and focal groups' population means differed (focal group N(−1,1) and reference group N(0,1)). In the second condition, the reference and focal groups' population standard deviations differed (focal group N(0,0.5) and reference group N(0,1)).

Test Length: Test lengths were set at 20, 40, and 80 items. Test lengths change between 20 and 80 items in most studies. In the literature, test lengths were defined as "short" for 20 items, "moderate" for 40 items, and "long" for 80 items. Moderate length (40 items) was preferred in many studies (Jodion & Gierl, 2001; Narayanan & Swaminathan, 1994; Rogers & Swaminathan, 1993). In Turkey, examination lengths range between 20 and 80.

Item Contamination (Proportion of Items with DIF): The proportion of items with DIF is another factor that affects the performances of DIF detection methods (Clauzer et al., 1993). When longer tests are used, more reliable scores are produced, and thus, more reliable trait estimation occurs. In addition, increasing the proportion of items with DIF produces less reliable trait estimation, and consequently, the power of the DIF detection methods decreases (Narayanan & Swaminathan, 1994). A high proportion of DIF items causes lower validity and decreases the power of DIF detection methods (Jodoin & Gierl, 2001). In this study, to test the effect of the proportion of DIF items, three levels were considered. Tests were simulated with 0%, 5%, and 10% of the items showing DIF.

Underlying Model Type: Studies have reported that underlying models influence the power of DIF detection methods (Cohen, Kim, & Wollack, 1996; Finch, 2005; Finch & French, 2007). The underlying models 2PL and 3PL were used to generate data in this study. The 3PL model was used to investigate the impact of pseudo-guessing parameters for the performance of DIF detection methods. The 2PL model was used to investigate the performances of methods in situations where pseudo-guessing was not possible.

Data Generation

IRT-LAB (Penfield, 2003) software was used to generate data based on the 2PL and 3PL models. In the generation of non-DIF items, the parameters from real testing conditions were used, i.e., from the Graduate Management Admission Test (Clauser et al., 1993). The a parameter of the items ranged between 0.29–1.40, the b parameter, between −2.95–2.12. In 3 PLM, the c parameter of all the items was selected as 0.20. In the generation of DIF items, the b parameter of the items was generated at approximately 0 (0.24–0.35), and the a parameter was generated at approximately 1 (0.78–1.11). In simulation studies, previous researchers were generally interested in moderate levels (Level B)

of DIF (Narayanan & Swaminathan, 1994; Rogers & Swaminathan, 1993), and some researchers indicated that in real-life situations, higher levels of DIF (level C) are uncommon (Linn, 1993). Some studies investigated DIF items on a large-scale test in Turkey. The results of these studies found some DIF items at C level (Bakan Kayalcıoğlu & Kelecioğlu; 2011; Doğan & Öğretmen, 2008), but generally, DIF items had negligible and moderate levels of DIF (Gök, Kelecioğlu, & Doğan, 2010; Karakaya, 2012; Karakaya & Kutlu, 2012). Hence parameter differences between focal and reference groups for DIF items were taken as 0.75. In DIF studies based on simulation, 100 replications are sufficient to reach consistent results (Kim, 2010). In the present study, 100 separate replications were conducted. According to NCME (2009) standards, 100 replications are sufficient to obtain consistent results, and the use of 100 replications is common (Kim, 2010). There is no limitation for the number of possible replications. Here evaluation indexes obtained by the number of replications were used to compare the methods. An appropriate replication number is required for consistent results. In the literature, most researchers used 100 replications (Ankenmann et al., 1996; Dainis, 2008; DeMars, 2009; Erdem Keklik, 2012; Fukuhara, 2009; Kim, 2010; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993), but many research studies have used more than 100 replications (Finch & French, 2007; Güler & Penfiled, 2009; Li, 2012; Li et al., 2012; Uttara & Millsap, 1994). According to Diaz-Emparanza (1996), the replication number should be as high as possible to reduce error.

In Turkey, research results revealed that non-uniform DIF is relatively rare compared to uniform DIF in actual testing practices (Bakan Kalaycıoğlu & Kelecioğlu, 2011; Doğan & Öğretmen, 2008). However, uniform DIF is most common, and non-uniform DIF relatively rare in test practices (Camilli & Shepard, 1994; Finch & French, 2007). Most DIF detection methods have the power to detect uniform DIF, but they are less powerful in detecting non-uniform DIF (Lopez, 2012). This study focuses only on uniform DIF. Thus, uniform DIF items were generated.

DIF analyses were conducted across 24 data sets, which combined two different sample sizes, two different trait distributions, three different levels of DIF contamination, and two different models (2x2x3x2). In this study, EZDIF (Waller, 1998) was used for the MH analyses. The SIBTEST and crossed SIBTEST analyses were obtained by using the SIBTEST software (Stout & Roussos, 1995), and IRTLRDIF was used for IRT-LR analyses (Thissen, 2001). One hundred replications were generated.

Evaluation Criteria

One evaluation criterion for DIF detection methods is Type I error rate. Type I error rates that are less than or equal to .05 indicate that the error of DIF detection methods is low. An empirical Type I error rate greater than the nominal alpha value (.05) is considered an inflated error. A second evaluation criterion is the power. Power values that are more than or equal to .80 show that the methods' power is sufficient. The criteria presented and used in this study are widely used in the literature. In this study, the data generation model was accepted as a condition. Results were analyzed using analysis of variance (ANOVA) to facilitate interpretation. Separate analyses were conducted for the different study criteria of Type I error and power. To compare the methods' performances (MH, IRT-LR, and SIBTEST), an ANOVA was conducted for each criterion. Then, to determine the effects of the study manipulations, separate analyses were conducted for each procedure. A full factorial ANOVA model was conducted including implementation, sample size, trait distribution, test length, ratio of items with DIF, and model type. However, the analysis could not be run because the error degrees of freedom were zero (Lopez, 2012). Instead, as a model for main effects, 2-, 3-, and 4-way interactions were run. Due to the large number of significance tests performed, a Bonferroni correction was used to control the family-wise error rate (resulting significance level, .002). Additionally, the eta-squared value for each term was reported to illustrate its effect on performance. Post-hoc analyses were run to test the significant ANOVA results.

Results

Type I Error

ANOVA results showed a significant difference among DIF detection methods ($F_{(2,213)} = 11.655$, $p = .000$). According to post-hoc test results, the Type I error of the SIBTEST method (.176) was significantly higher than those of the MH (.129) and IRT-LR (.111) methods.

All main effects were significant, while some of the interaction effects were significant for models. Sample size provided significant results for the MH method. In the case of unequal sample sizes for focal and reference groups, Type I error was found to be

decreasing in the MH method. Ability distribution had significant effect on the SIBTEST methods. Type I error was significantly lower under the condition in which the reference and focal groups' population standard deviations differed. In the SIBTEST and MH methods, test length caused significant differences. In the SIBTEST method, test length caused a decrease in the Type I error. In the MH method, there were significant differences between tests with 20 items and those with 40–80 items. Tests with 20 items had lower Type I error, but non-significant differences were found between tests with 40 and 80 items. The ratio of DIF items in the test affected all methods' Type I error. Increments of the ratio of DIF items caused increases in Type I error in the IRT-LR and SIBTEST methods. In the MH procedure, the lowest Type I error was detected in the test with 5% DIF items. Post-hoc results were significant in all pairwise comparisons. Model type had a significant effect on the Type I error rate of IRT-LR and SIBTEST. The Type I error rate was higher in the 3PLM for SIBTEST and 2PLM for IRT-LR.

Power

ANOVA results showed significant differences among DIF detection methods ($F_{(2,141)} = 7.573$, $p = .000$). According to post-hoc test results, the powers of the MH (.999) and SIBTEST (.995) methods were significantly higher than the IRT-LR (.973) method.

Separate ANOVA analyses were conducted to detect the power of methods under different conditions. Results showed that the conditions and the interaction of conditions did not affect the power of the SIBTEST and MH methods. But the underlying model had significant effect on the IRT-LR power rate ($F_{(1,48)} = 720.474$, $p = .001$, $\eta^2 = .997$). The power of the IRT-LR was higher in the 2PLM. The interaction effect of sample size*model type*trait distribution was effective in the power of the IRT-LR ($F_{(2,72)} = 475.0$, $p = .002$, $\eta^2 = .996$). Thus, the power rate of the IRT-LR was lower under the condition in which 3PLM with equal sample sizes and the reference and focal groups' population means were equal (standard deviations differed) than in other conditions.

Discussion

The proportion of DIF items and model types are effective factors for IRT-LR. Most of the research has had the same results: an increase in the ratio of DIF items causes a higher rate of Type I error (Finch, 2005; Stark, Chernyhenko, & Drasgow, 2006; Wang & Yeh, 2003). In 3PLM, the Type I error rate of the IRT-LR was lower. Research results on this issue were complex. Some results (Cohen et al., 1996) showed that the Type I error rate in the 3 PLM was lower, but some (Finch, 2005) revealed it to be lower in the 2 PLM. Finch (2005) compared the focus and reference groups in a smaller sample (600 and 1000) when the proportion of the DIF items was 15%. Finch's results resemble those of the present study. In all DIF detection methods (IRT-LR, SIBTEST, and MH), Type I error decreased, and the power rate increased when the sample sizes for focal and reference groups changed.

Test length had a more powerful effect on the SIBTEST than on other methods. Increments in test length caused decreases in the Type I error rate. The power rate of the SIBTEST was 0.99 in all conditions. Thus, the differences in test length did not affect the SIBTEST's power.

According to the trait distribution differences, the highest Type I error and the lowest power rate were seen in the SIBTEST. In both trait distribution conditions, the lowest Type I error was found in the IRT-LR method, and the highest power was found in the MH procedure. The Type I error rate of the SIBTEST and MH was low, and the differentiation in trait distribution caused an increase in the Type I error rate of both methods (Roussos & Stout, 1996); SIBTEST was more powerful when the reference and focal groups' population means differed (standard deviations were equal) in the detection of non-uniform DIF (Naranayan & Swaminhatan, 1996). Pei and Li's results (2010) differed from the present study's. The variance differences between focal and reference groups increased the Type I error rate of MH more than that of the SIBTEST. Similarly to the present study, they found that variance differences had minimal effect on the IRT-LR method.

Sample size ratio for focal and reference groups was effective for the MH procedure. Type I error was lower in unequal sample sizes. This finding was supported by previous research (Kristjansson et al., 2005). The most powerful method was MH (0.99–1.00), and the Type I error rate of the SIBTEST was found to be highest for this study.

Future research will compare the different DIF detection methods in different samples. In addition to this, future research will investigate these conditions by eliminating DIF contamination. The present study's results have provided information for uniform DIF. Future research will detect non-uniform DIF in similar conditions.

References/Kaynakça

Ackerman, T. A., & Evans, J. A. (1992, April). *An investigation of the relationship between reliability, power, and the Type I error rate of the Mantel-Haenszel and simultaneous item bias detection procedures*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Agresti, A. (1984). *Analysis of ordinal categorical data*. New York: John Wiley & Sons.

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1996, April). *An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Atalay, K., Gök, B., Kelecioğlu, H. ve Arsan, N. (2012). Değişen madde fonksiyonunun belirlenmesinde farklı yöntemlerin kullanılması: Bir simülasyon çalışması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *43*, 270-281.

Bakan Kalaycıoğlu, D. ve Kelecioğlu, H. (2011). Öğrenci Seçme Sınavı'nın madde yanlılığı açısından incelenmesi. *Eğitim ve Bilim, 36*(161), 3-12.

Bekçi, B. (2007). *Ortaöğretim kurumları öğrenci seçme sınavının değişen madde fonksiyonlarının cinsiyete ve okul türüne göre incelenmesi* (Yüksek lisans tezi, Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara). http://tez2.yok.gov.tr/ adresinden edinilmiştir.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., Vol. 4, pp. 221-256). Westport: American Council on Education & Praeger Publishers.

Camili, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* London: Sage.

Cheng, C. M. (2005). *A study on differential item functioning of the basic mathematical competence test for junior high schools in Taiwian* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database. (UMI No. 3189625)

Clauser, E. B., Mazor, K., & Hambleton, K. R. (1993). The effects of purification of the matching criterion on the identification of DIF using theMantel-Haenszel procedure. *Applied Measurement in Education*, *6*(4), 269–279. doi:10.1207/s15324818ame0604_2

Cohen, A. S., Kim, S., & Wollack, J. A. (1996). An investigation of likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement*, *20*(1), 15-26. doi: 10.1177/014662169602000102

Çepni, Z. (2011). *Değişen madde fonksiyonlarının sıbtest, Mantel Haenszel, lojistik regresyon ve madde tepki kuramı yöntemleriyle incelenmesi* (Doktora tezi, Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara). http://tez2.yok.gov.tr/ adresinden edinilmiştir.

Dainis, A. M. (2008). *Methods for identifying differential item and test functioning: an investigation of type I error rates and power* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database. (UMI No. 3323367)

DeMars, C. E. (2009). Modification of the Mantel-Haenszel and logistic regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of Educational and Behavioral Statistics, 34,* 149-170. doi:10.3102/1076998607313923

Diaz-Emparanza, I. (1996). Selecting the Number of Replications in a Simulation Study. Economics Working Paper Archieve University of Washington, Ref: ewp-em/9612006, EconWPA. Retrived from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1582

Doğan, N. ve Öğretmen, T. (2008). Değişen madde fonksiyonunu belirlemede Mantel- Haenszel, Ki-kare ve lojistik regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim, 33*, 100-112.

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

Erdem Keklik, D. (2012). *İki kategorili maddelerde tek biçimli değişen madde fonksiyonu belirleme tekniklerinin karşılaştırılması: bir simülasyon çalışması* (Doktora Tezi, Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü). http://tez2.yok.gov.tr/ adresinden edinilmiştir.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*, 278-295. doi: 10.1177/0146621605275728

Finch, H. W., & French, F. B. (2007). Detection of crossing differential item functioning: A comparison of four methods. *Educational and Psychological Measurement, 67*(4), 565-582. doi: 10.1177/0013164406296975

Fukuhara, H. (2009). *A differential item functioning modelfor testlet-based items using a bi-factor multidimensional item response theory model: A Bayesian approach* (Doctoral Dissertation). Retrieved from: http://diginole.lib.fsu.edu/cgi/viewcontent.cgi?article=1573&context=etd

Gök, B., Kelecioğlu, H. ve Doğan, N. (2010). Değişen madde fonksiyonunu belirlemede Mantel-Haenszel ve lojistik regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim, 35,* 156, 3-16.

Güler, N., & Penfield, R. D. (2009). A comparison of the logistic regression and contingency table methods for the simultaneous detection of uniform and nonuniform DIF. *Journal of Educational Measurement, 46,* 314-329. doi: 10.1111/j.1745-3984.2009.00083.x

Hidalgo, M. D., & Lopez-Pina, J. A. (2004). Differential item functioning detection an effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*(6), 903-915. doi: 10.1177/0013164403261769

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

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum.

Jodoin, G. M., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Educatıon, 14*(4), 329–349. doi:10.1207/S15324818AME1404_2

Karakaya, İ. (2012). Seviye Belirleme Sınavındaki Fen ve Teknoloji ile Matematik alt testlerinin madde yanlılığı açısından incelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri, 12*, 215-229.

Karakaya, İ. ve Kutlu Ö. (2012). Seviye Belirleme Sınavındaki Türkçe alt testlerinin madde yanlılığı açısından incelenmesi. *Eğitim ve Bilim, 37*, 165, 348-362.

Kim, J. (2010). *Controlling type 1 error rate in evaluating differential item functioning for four DIF methods: Use of three procedures for adjustment of multiple item testing* (Doctoral Dissertation). Retrieved from http://scholarworks.gsu.edu/eps_diss/67/

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting DIF in ordered response items. *Educational and Psychological Measurement, 65*, 935–953. doi: 10.1177/0013164405275668

Li, H. H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika*, *61*, 647-677.

Li, Y. (2012). *Item discrimination and type I error rates in DIF detection using the Mantel-Haenszel and logistic regression procedure (Doctoral Dissertation).* Available from ProQuest Dissertations and Theses database. (UMI No. 3540134)

Li, Y., Brooks, G. P., & Johanson, G. A. (2012). Item discrimination and type I error in the detection of differential item functioning. *Educational and Psychological Measurement, 72*(5), 847-861.

Linn, R. L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 349-364). Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Lopez, G. E. (2012). *Detection and classification of DIF types using parametric and nonparametric methods: A comparison of the IRT-Likelihood Ratio Test, Crossing-SIBTEST, and logistic regression procedures* (Doctoral Dissertation). Retrieved from http://scholarcommons.usf.edu/etd/4131/

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential. *Applied Psychological Measurement, 18*(4), 315-328. doi: 10.1177/014662169401800403

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show non-uniform DIF. *Applied Psychological Measurement, 20*(3), 257–274. doi: 10.1177/014662169602000306

Pei, L. K., & Li, J. (2010). Effects of unequal ability variances on the performance of logistic regression, Mantel-Haenszel, SIBTEST IRT, and IRT Likelihood ratio for DIF detection. *Applied Psychological Measurement, 36*(4), 453–456. doi:10.1177/0146621610367789

Penfield, R. D. (2003). Application of the BreslowDay test of trend in odds ratio heterogeneity to the detection of nonuniform DIF. *Alberta Journal of Educational Research*, *49*, 231-243.

Potenza, T. M., & Dorans, J. N. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*(1), 23–37. doi: 10.1177/014662169501900104

Prieto, P., Barbero, M. I., & San Luis, C. (1997). Identification of nonuniform DIF: A comparison of Mantel-Haenszel and IRT analysis procedure. *Educational and Psychological Measurement, 57*(4), 559-568.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*, 105-116. doi: 10.1177/014662169301700201

Roussos, L. A. (1992). *Hierarchical agglomerative clustering computer program user's manual* (Unpublished manuscript). University of Illinois at Urbana-Champaign.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33*, 215-230.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/ DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159–194.

Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6*, 317-375. doi: 10.2307/1164616

Shepard, L. A., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9*, 93–128. doi: 10.2307/1164716

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with CFA and IRT: Toward a unified strategy. *Journal of Applied Psychology, 91*, 1292-1306.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods, 11*, 402–415. doi: 10.1037/1082-989X.11.4.402

Stout, W. F., & Roussos, L. A. (1995). *SIBTEST user's manual [Computer program manual]* (2nd ed.). Urbana-Champaign: University of Illinois, Department of Statistics.

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

Thissen, D. (2001). *IRTLRDIF (Version 2.0b)* [Computer software]. Chapel Hill: L. L. Thurstone Psychometric Laboratory, University of North Carolina.

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

Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement, 18*, 15-25.

Vaughn, B. K., & Wang, Q. (2010). DIF trees: Using classification trees to detect differential item functioning. *Educational and Psychological Measurement, 70*(6), 941-952. doi: 10.1177/0013164410379326

Waller, N. G. (1998). EZDIF: Detection of uniform and non-uniform differential item functioning with the mantel-haenszel and logistic regression procedures. *Applied Psychological Measurement*, *22*(4), 391-391. doi: 10.1177/014662169802200409

Wang, W. C., & Yeh, Y. L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*, 479-498. doi: 10.1177/0146621603259902

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Hillsdale,NJ: Erlbaum

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*(3), 185-197. doi: 10.3102/10769986015003185