International Journal of Progressive Education, Volume 15 Number 4, 2019

Investigating Differential Item Functioning in DINA Model

Seçil Ömür Sünbül [i]
Mersin University

Abstract

In this study, it is aimed to investigate the effects of various factors on the performance of the methods used in the determination of differential item functioning (DIF) in the DINA model included in the Cognitive Diagnosis Models. The current study is limited with Logistic Regression and Wald test methods which were used to determine the differential item functioning in DINA model. The Type I error and power rates of these methods in certain conditions were investigated to evaluate their performances. In the simulation study for the Type I error rates, four variables were manipulated: sample sizes, the number of attributes, correlations between attributes and reference group s and g parameter values. In the determination of the power rates of the methods, additionally, the variables that were manipulated in the Type I error study, DIF sizes and percentages of DIF items were manipulated, too. As a result, it was observed that especially in all cases where reference group' s and g parameter values are low, both methods yielded a good control of Type I error rates. In addition, according to the results, it was observed that both DIF size and sample size affect the power rates of both methods.

Keywords: DINA model, differential item functioning, Wald test, logistic regression

DOI: 10.29329/ijpe.2019.203.13

---

[i] **Seçil Ömür Sünbül,** Assist. Prof. Dr., Mersin University, Measurement and Evaluation in Education, Mersin/Turkey.

Correspondence: secilomur@gmail.com

INTRODUCTION

In recent years, Cognitive Diagnostic Models (CDMs) have been widely used in education and psychology. CDMs are models that provide information about the strengths and weaknesses of individuals in specific areas. CDMs are latent variable models developed primarily for assessing student mastery and non-mastery on a set of finer-grained skills (de la Torre, 2011). The results obtained from CDMs provides detailed feedback to the examinees or teacher, so they can make inferences about examinees' mastery of different cognitive skills.

Most CDMs applications require the construction of a Q-matrix (Embretson, 1984; Tatsuoka, 1985, de la Torre, 2009). The relationship between items and attributes is specified in the Q-matrix, which is a matrix with *j* rows and *k* columns of ones and zeros. q_{jk} is an element of Q matrix for *j* items and *k* attributes indicates whether mastery of attribute *k* is required by item *j*. q_{jk} = 1, if item *j* requires attribute *k*, and 0 otherwise.

When the related literature was investigated, several CDMs have been developed for assessing examinees' mastery or non-mastery of a set of cognitive attributes (Haertel, 1989; Dibello et al., 1995; Junker & Sijtsma, 2001; Hartz, 2002; de la Torre & Douglas, 2004; Templin & Henson, 2006; Templin, Henson & Douglas, 2006; Henson, Templin, & Willse, 2009). In this study, the deterministic, inputs, noisy "and" gate (DINA) model, which is one of the most widely used non-complementary models developed by Haertel (1989), was used. DINA model assumes examinees must have mastered a set of attributes required by an item in order to answer the item correctly. The DINA is a simple model that is easily estimated and the item response function is given by

$$P(X_{ij} = 1|\alpha_{ij}) = (1 - s_j)^{\eta_{ij}} g_j^{(1-\eta_{ij})} \quad (1)$$

where *P* denotes the probability of solving the item when examinees possess all of the required skills. *Xij* denotes the response of an examinee *i* to item *j*, where *Xij* = 1 is the correct response (*Xij* = 0 otherwise). *gj* denotes guessing parameter and *sj* denotes slipping parameters for the jth item. The slip parameter is interpreted as the probability that examinee who possesses all the required attributes for an item answers the item incorrectly (de la Torre ve Lee, 2010). The guessing parameter is the probability that examinee who lacks at least one of the required attributes for an item answers the item correctly (de la Torre ve Lee, 2010). When a slip parameter is low, the examinee has a higher probability of answering the item correctly. ηij is the deterministic latent response and it is given by

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} \quad (2)$$

qjk = assumes 1 or 0, αik = 1 or 0, represents if examinee i mastered attribute k. If ηij = 1, represents examinee *i* possesses all the attributes required for item *j*, and ηij = 0 represents examinee *i* lacks at least one of the attributes required for item *j*.

Differential Item Functioning

Analysis for detecting Differential Item Functioning (DIF) has been increasingly applied in test fairness studies. DIF occurs when individuals at the same ability level but in different subgroups differ in their probability of answering an item correctly (Zumbo, 1999; Hambleton, Swaminathan, & Rogers, 1991). In the DIF analysis, the group which is thought to be disadvantagous is called focal group while the advantageous group compared with the performance of this group is called the reference group. DIF can occur in two different ways and the first is the uniform DIF. Uniform DIF indicates that the difference in the probability of answering an item correctly is consistent at all levels of ability. The second is the non-uniform DIF and implies that the difference in the probability of responding correctly is different for all ability level range (Camilli & Shepard, 1994; Zumbo, 1999). In order to determine DIF many methods have been developed within the context of both Classical Test Theory (CTT) and Item Response Theory (IRT). While methods such as Mantel-Haenszel (MH),

Logistic Regression (LR) and the simultaneous item bias test (SIBTEST) are investigated under CCT, the methods such as likelihood ratio test, Lord $\chi 2$ and Raju's area measurements are investigated under IRT (Raju, 1988; Hambleton, Swaminathan & Rogers, 1991; Rogers & Swaminathan, 1993; Camilli & Shepard, 1994; Osterlind, 1983).

In CDMs, DIF occurs when individuals with different groups but with the same attribute mastery profile differ in their probability of responding correctly to the item. For the DINA model, DIF occurs when different estimates obtained for the slip and guess parameters for the individuals in the focal and reference groups. Uniform DIF occurs in item *j* when Δ_{sj} and Δ_{gj} have the same signs (Hou et al., 2014);

$$\begin{aligned} \Delta_{sj} &> 0 \quad \text{or } s_{Fj} - s_{Rj} < 0 \\ \Delta_{gj} &> 0 \quad \text{or } g_{Fj} - g_{Rj} > 0 \end{aligned} \tag{3}$$

$$\begin{aligned} \Delta_{sj} &< 0 \quad \text{or } s_{Fj} - s_{Rj} > 0 \\ \Delta_{gj} &< 0 \quad \text{or } g_{Fj} - g_{Rj} < 0 \end{aligned} \tag{4}$$

When Equation 3 is investigated, uniform DIF in item *j* occurs when the slip parameter in the focal group is smaller than the slip parameter in the reference group and the guessing parameter in the focal group is larger than the guessing parameter in the reference group. When Equation 4 is investigated, uniform DIF in item *j* occurs when the slip parameter in the focal group is larger than the slip parameter in the reference group, and the guessing parameter is smaller than the guessing parameter in the reference group.

Nonuniform DIF occurs in item j when $\Delta_{sj}$and Δ_{gj} have different signs (Hou et al., 2014);

$$\begin{aligned} \Delta_{sj} &> 0 \quad \text{or } s_{Fj} - s_{Rj} < 0 \\ \Delta_{gj} &< 0 \quad \text{or } g_{Fj} - g_{Rj} < 0 \end{aligned} \tag{5}$$

$$\begin{aligned} \Delta_{sj} &< 0 \quad \text{or } s_{Fj} - s_{Rj} > 0 \\ \Delta_{gj} &> 0 \quad \text{or } g_{Fj} - g_{Rj} > 0 \end{aligned} \tag{6}$$

When Equation 5 and 6 are investigated, nonuniform DIF in item j occurs when both the slip and guess parameters in the focal group are smaller than the reference group or when both the slip and guess parameters in the focal group are larger than the reference group.

When the relevant literature is investigated, it was observed that there is a limited study on DIF in CDM framework (Zhang, 2006; Li, 2008; Hou et al., 2014; Li and Wang, 2015). Zhang (2006) studied DIF in the DINA model using Mantel-Haenszel and SIBTEST methods in both real and simulation data. Four variables were manipulated in the simulation study: sample sizes, types of DIF, levels of DIF amount, and correlations between skill attributes. It was observed that attribute pattern matching had lower Type I error rates and higher power rates than the traditional total test score matching under the comparable test conditions. Li (2008) used a modified higher order DINA model to investigate DIF and differential attribute functioning (DAF). Five factors were manipulated in the simulation study: Q-matrix structure, attribute discrimination parameters, sample size, ability distribution difference, scenarios of DIF and DAF combination. For DIF detection, the model-based method was also compared with the MH method using a total score as the matching criterion and an attribute profile as the matching criterion. It was observed that the recovery of item parameters was generally better than the recovery of attribute parameters. In addition, it was observed that, model-based method had better Type I error rates and had higher power rates than the Mantel–Haenszel. Hou

et al. (2014) used a DINA model to investigate the effectiveness of the Wald test in detecting DIF. They compare the Wald test with both Mantel–Haenszel and SIBTEST procedures. The sample size, reference item parameters, DIF size, and DIF type were manipulated in the simulation study. They found that the performance of the Wald test was not affected by the proportion of DIF items in the test and both for small and large sample sizes the Wald test has Type I error rates close to the nominal level. Li and Wang (2015), developed a general CDM-based method for DIF assessment. They were compared performance of LCDM-DIF and Wald methods. When two groups were investigated, they found that when tests were clean, both methods yielded a good control of Type I error rates When all items were DIF, the power rates of the LCDM-DIF method were higher than the power rates of the Wald method with two groups. When three groups were investigated, they found that, the LCDM-DIF method had a good control of Type I error rates under all conditions, however, even if the tests were clean, the Type I error rates of the Wald test were higher.

In this study, it is aimed to investigate the effect of various factors on the performance of the methods used in the determination of differential item functioning in the DINA model. When the related literature is investigated, it is considered that this study will contribute to the field since it has investigated different factors and factors' levels.

METHOD

Simulation Design

Sample Size: Zhang (2006) used the equal sample sizes of 400 and 800 for focal and reference groups in his simulation study. Other than this study, Li (2008), Hou et al. (2014), Li and Wang (2015) simulated equal sample sizes (500 and 1000) for focal and reference groups. In this study, three sample sizes, 500, 1000 and 2000 were used for each group, in order to compare the results of the current study with the related literature.

Correlation Between Attributes: When the studies about CDM and DIF are investigated, it was observed that in some studies the correlation between the attributes were kept constant (e.g. Hou et al. 2014; Li and Wang, 2015), and in some studies this factor is manipulated in various ways (e.g. Zhang, 2006). In this study, correlations between attributes were manipulated as low (0.2), medium (0.5) and high (0.8).

Number of Attribute and Item: Zhang (2006) and Li (2008) used a test which contains 5 attributes and 25 items in their study. However, Hou et al. (2014) used a test of 5 attributes and 30 items, Li and Wang (2015) used 5 attributes, 30 and 50 items in their studies. In this study, the number of items was fixed to 30, and the number of attributes was manipulated as 4 and 5. Q matrices were generated according to the number of attributes. Q matrices were generated in such a way that a maximum of three attributes is observed in an item.The generated Q matrices are shown in Table 1.

s and g parameter Values of the Reference group: In this study, s and g parameter values of the reference group were manipulated as three levels: 0.1, 0.2 and 0.3.

DIF type and size: Magnitude of DIF varies according to the models used in the studies. DIF size levels of this study is similar to Zhang's (2006)**,** and Hou et al.'s (2014) studies such as (Δ_{sj}or Δ_{gj} = .05) for small DIF size and (Δ_{sj}or Δ_{gj} = .10) for large DIF size.

Percentage of DIF Items: Related studies indicate that the percentage of DIF items in an overall test, affects the performance of DIF detection methods (Zhang, 2007; Hou et al., 2014). In this study, the percentage of DIF items in the test was manipulated as 10% and 20%.

DIF Detection Methods: In this study, DIF was limited by using LR and Wald methods.

Table 1: Q-Matrices for the Simulated Data

Q1				
Attribute				
Item	1	2	3	4
1	1	0	0	0
2	1	0	0	0
3	1	0	0	0
4	0	1	0	0
5	0	1	0	0
6	0	1	0	0
7	0	0	1	0
8	0	0	1	0
9	0	0	0	1
10	0	0	0	1
11	1	1	0	0
12	1	1	0	0
13	1	0	1	0
14	1	0	1	0
15	1	0	0	1
16	1	0	0	1
17	0	1	1	0
18	0	1	1	0
19	0	1	0	1
20	0	1	0	1
21	0	0	1	1
22	1	1	1	0
23	1	1	1	0
24	1	1	0	1
25	1	1	0	1
26	1	0	1	1
27	1	0	1	1
28	0	1	1	1
29	0	1	1	1
30	1	1	1	0

Q2					
Attribute					
Item	1	2	3	4	5
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1
6	1	0	0	0	0
7	0	1	0	0	0
8	0	0	1	0	0
9	0	0	0	1	0
10	0	0	0	0	1
11	1	1	0	0	0
12	1	0	1	0	0
13	1	0	0	1	0
14	1	0	0	0	1
15	0	1	1	0	0
16	0	1	0	1	0
17	0	1	0	0	1
18	0	0	1	1	0
19	0	0	1	0	1
20	0	0	0	1	1
21	1	1	1	0	0
22	1	1	0	1	0
23	1	1	0	0	1
24	1	0	1	1	0
25	1	0	1	0	1
26	1	0	0	1	1
27	0	1	1	1	0
28	0	1	1	0	1
29	0	1	0	1	1
30	0	0	1	1	1

Data Generation and Analysis

In this study, data were generated according to DINA model. To generate data, the number of items was set to 30, and Q matrices were formed with the number of attributes which is 4 and 5. s and g parameter values of the reference and focal groups were manipulated into three levels such as 0.1, 0.2 and 0.3. To determine the Type I error rates of the methods, three sample sizes (500, 1000 and 2000), two number of attributes (4 and 5), three correlation between attribute levels (0.2, 0.5 and 0.8) and three reference group item parameters (0.1, 0.2 and 0.3) were manipulated. The sample sizes were formed as equal in the reference and focal groups. In the determination of the power rates of the methods, in addition to the variables that were manipulated in the Type I error study, two DIF sizes (0.05 and 0.1) and two DIF item percentages (10% and 20%) were manipulated, too. The slip and the guessing parameter values for the focal group were manipulated according to the DIF size. DIF items were generated according to the percentage of DIF. While forming DIF items, the number of attributes**,** which is required for that item**,** was also considered for balancing. For example, 3 DIF items were generated for 10% condition. One of them was requiring 1 attribute, the other one was requiring 2 attributes and the last one was requiring 3 attributes. For 20% condition, the number of items were doubled for each requirement case. In this study, only uniform DIF type has been investigated and the summary of DIF conditions is shown in Table 2. 100 replications were conducted for each crossing

condition. R 3.5.1 was used as the programming language and *CDM, GDINA* and *difR* package were used for data generation and data analysis.

For Type I error rates, the false positive rates for items, which were detected incorrectly as DIF items, were reported over 100 replications. However, in power study the true positive rates were obtained for determining items, which perform differently on different groups of examinees.

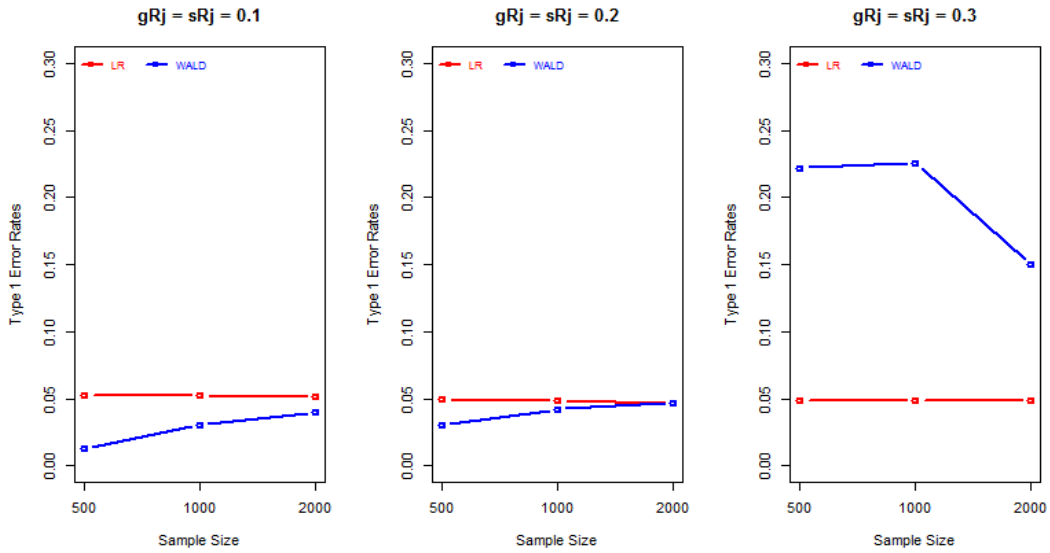
Table 2: Summary of DIF Conditions

DIF Type	DIF Size	Δ_{sj} (s_{Fj} - s_{Rj})	Δ_{gj} (g_{Fj} - g_{Rj})
Non- DIF	-	0	0
Uniform	Small	+0.05	+0.05
		-0.05	-0.05
	Large	+0.1	+0.1
		-0.1	-0.1

FINDINGS

Type I Error Study

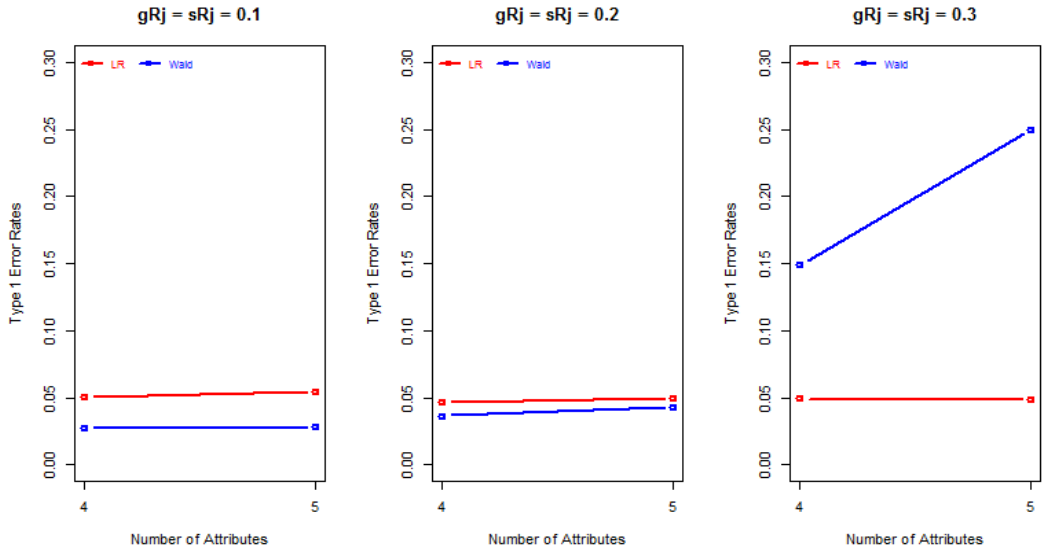
The results of the effects of various sample sizes on the Type I error rates of methods are shown in Graph 1. When Graph 1 was investigated, it was observed that the Type I error rates of the Logistic Regression method were not affected by the increase in sample size for all s and g parameter values. However, in cases where s and g parameter values were 0.1 and 0.2, the Type I error rates of the Wald test method were not effected with the increase in sample size. When the s and g parameter value was 0.3, it was observed that the Type I error rates of the Wald test method decreased dramatically with the increasing sample size. Also, when the s and g parameter values were 0.1 and 0.2, the Type I error rates of the Wald test method were lower than the Type I error rates of the LR method in all sample sizes. However, when the s and g parameter value was 0.3, the Type I error rates of Wald test were larger than the Type I error rates of the LR method.



Graph I: The Effect of the Sample Sizes on Type I Error Rates of Methods

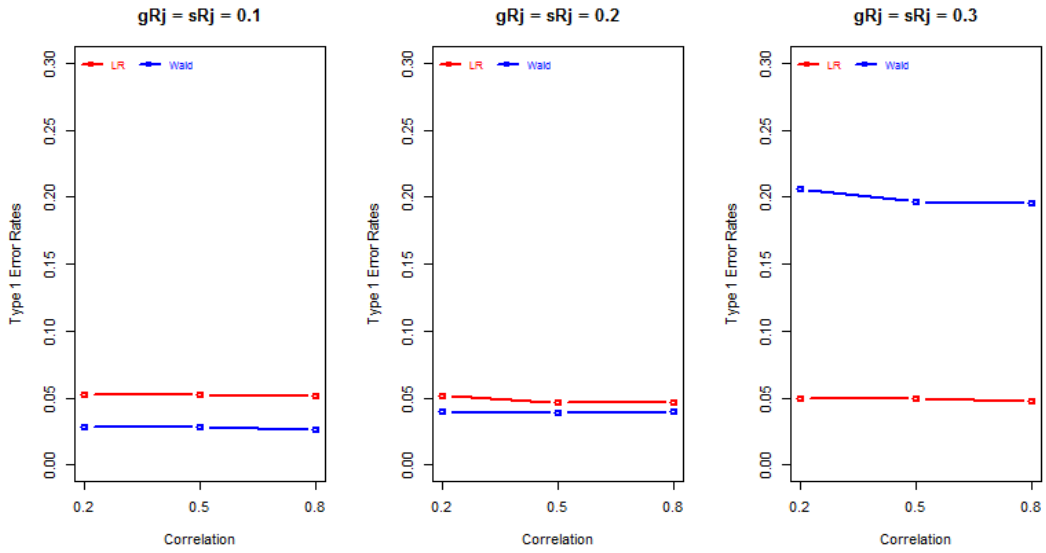
The results of the effects of the number of attributes on the Type I error rates of methods are shown in Graph 2. According to Graph 2, especially when the s and g parameter values were 0.1 and 0.2, it was observed that the Type I error rates of the methods did not change much with the increase in the number of attributes. However, when the s and g parameter value was 0.3, the Type I error rates of the Wald test method increased with the increase in the numbers of attributes. In the case where s

and g parameter values were 0.1 and 0.2 for both numbers of attributes (4 and 5), the Type I error rates of the Wald test method were smaller than the Type I error rates of the LR method.



Graph 2: The Effect of the Number of Attributes on Type I Error Rates of Methods

The results of the effects of the correlation between attributes on the Type I error rates of methods are shown in Graph 3. When Graph 3 was investigated, it was observed that the Type I error rates of the methods did not change much with the increase in correlation levels between the attributes. However, when the s and g parameter values were 0.1 and 0.2, the Type I error rates of the Wald test method were lower than the Type I error rates of the LR method for all correlation levels. In the case where s and g parameter value was 0.3, a significant increase was observed for the Type I error rates of the Wald test method in this case the Type I error rates of the Wald test method were higher than the Type I error rates of the LR method.



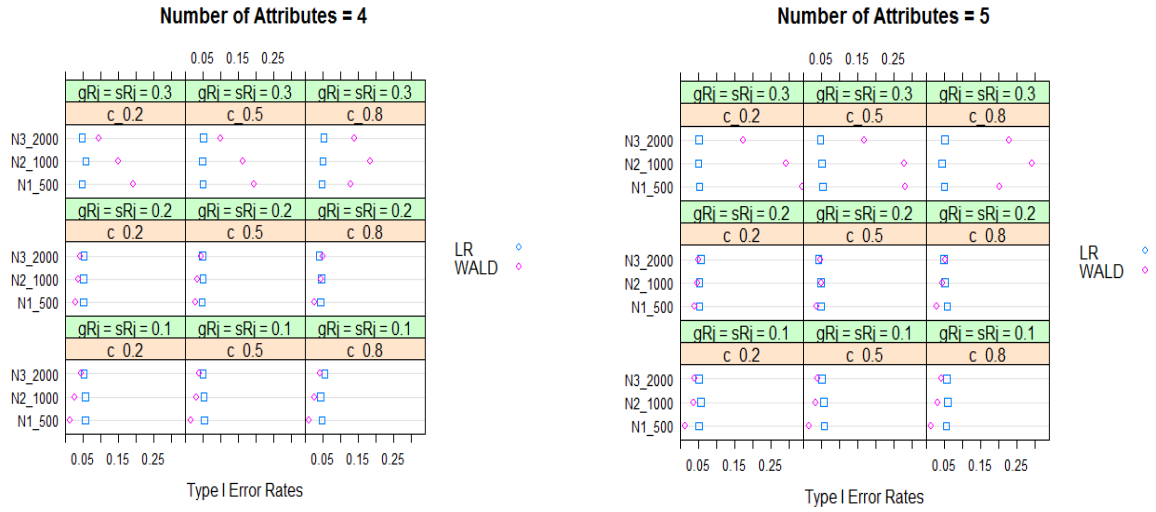
Grafik 3: The Effect of the Correlation between Attributes on Type I Error Rates of Methods

Results of the Type I error rates of the methods according to all the manipulated factors shown in Table 3 and Graph 4. When Graph 4 was investigated, it was observed that the s and g parameter values were effective for Type I error rates of the Wald test method. According to Graph 4, especially in cases where the s and g parameter values were 0.1 and 0.2, the Type I error rates of the Wald test

method were lower than the Type I error rates of the LR method for all conditions. In contrast, in the case where s and g parameter value was 0.3, it was observed that the Type I error rates of the LR method were lower than the Type I error rates of the Wald test method. When the s and g parameter value was 0.3, it was observed that the Type I error rates of the Wald test method increased with the increase in the number of attributes and the decrease in the sample size. In addition to that, the Type I error rates of the LR method did not change much with the increase in the sample size.

Tablo 3: Type I Error Rates

			DIF Detection Method					
			Sample Size					
			LR			Wald		
Reference Item Parameter Values	**Number of Attribute**	**Correlation**	***N*R = 500 *N*F = 500**	***N*R = 1,000 *N*F = 1,000**	***N*R = 2,000 *N*F = 2,000**	***N*R = 500 *N*F = 500**	***N*R = 1,000 *N*F = 1,000**	***N*R = 2,000 *N*F = 2,000**
$g_{Rj}= s_{Rj}= 0.1$	**4**	0.2	0.055	0.055	0.051	0.013	0.026	0.044
		0.5	0.053	0.051	0.049	0.016	0.030	0.038
		0.8	0.047	0.042	0.054	0.009	0.026	0.041
	5	0.2	0.050	0.056	0.050	0.013	0.037	0.038
		0.5	0.057	0.056	0.050	0.015	0.033	0.038
		0.8	0.055	0.058	0.056	0.012	0.031	0.042
$g_{Rj}= s_{Rj}= 0.2$	**4**	0.2	0.050	0.050	0.052	0.028	0.037	0.041
		0.5	0.047	0.049	0.048	0.027	0.033	0.045
		0.8	0.043	0.045	0.038	0.025	0.043	0.050
	5	0.2	0.051	0.051	0.055	0.039	0.046	0.050
		0.5	0.048	0.048	0.041	0.035	0.050	0.046
		0.8	0.057	0.050	0.048	0.027	0.045	0.051
$g_{Rj}= s_{Rj}= 0.3$	**4**	0.2	0.046	0.057	0.046	0.192	0.148	0.094
		0.5	0.049	0.048	0.050	0.195	0.162	0.098
		0.8	0.046	0.049	0.052	0.128	0.183	0.139
	5	0.2	0.051	0.049	0.050	0.336	0.291	0.174
		0.5	0.053	0.050	0.047	0.280	0.278	0.168
		0.8	0.049	0.042	0.050	0.201	0.291	0.229

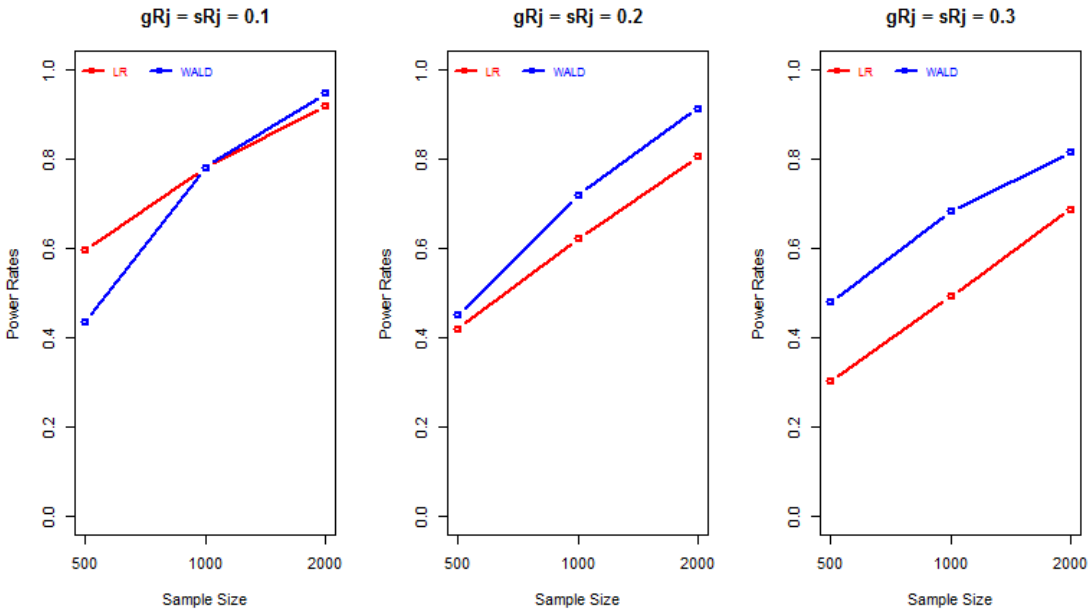


Graph 4: The Interaction Effect of Factors on Type I Error Rates of Methods

Power Study

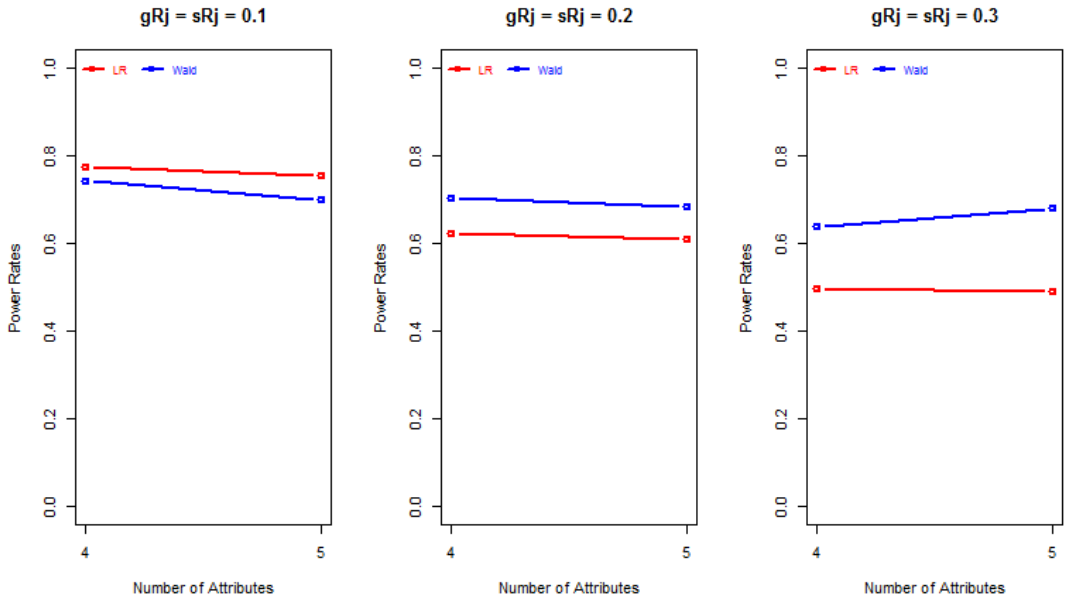
The results of various sample sizes on the power rates of the methods are shown in Graph 5. According to Graph 5, it was observed that, the power ratios of both methods were increasing with the increase in sample size. When the s and g parameter values increase, there is a decrease in the power rates of the LR method. For the condition in which s and g parameter value was 0.1 and the sample size was 500, it was observed that, the power rates of the Wald test method were lower than the power rates of the LR method. However, when the sample size was 1000, it was observed that the power

rates of the methods were similar, whereas when the sample size was 2000, it was observed that the power rates of the Wald test method were higher than the power rates of the LR method. Also, when the s and g parameter values were 0.2 and 0.3, it was observed that, the power rates of the Wald test were higher than the LR method for all sample sizes.



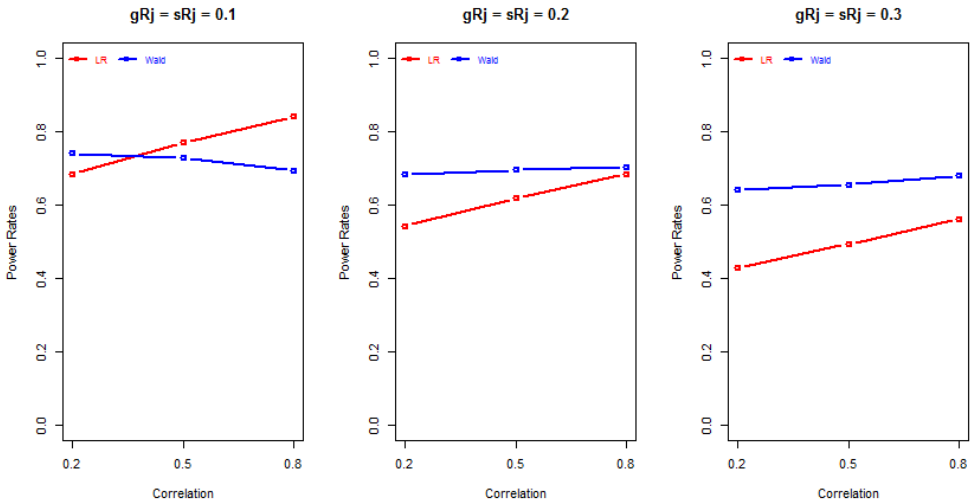
Graph 5: The Effect of the Sample Sizes on Power Rates of Methods

The results of the power rates of the methods with the different number of attributes were shown in Graph 6. According to Graph 6, it was observed that, the power rates of the Wald test method did not change much with the increase in the number of attributes, but the power rates of the LR method decreased. When the s and g parameter value was 0.1, it was observed that, the Type I error rates of the Wald test method was lower than the Type I error rates of the LR method for all number of attributes, but the s and g parameter values were 0.2 and 0.3, the Type I error rates of the Wald test method was higher than the Type I error rates of the LR method for all number of attributes.



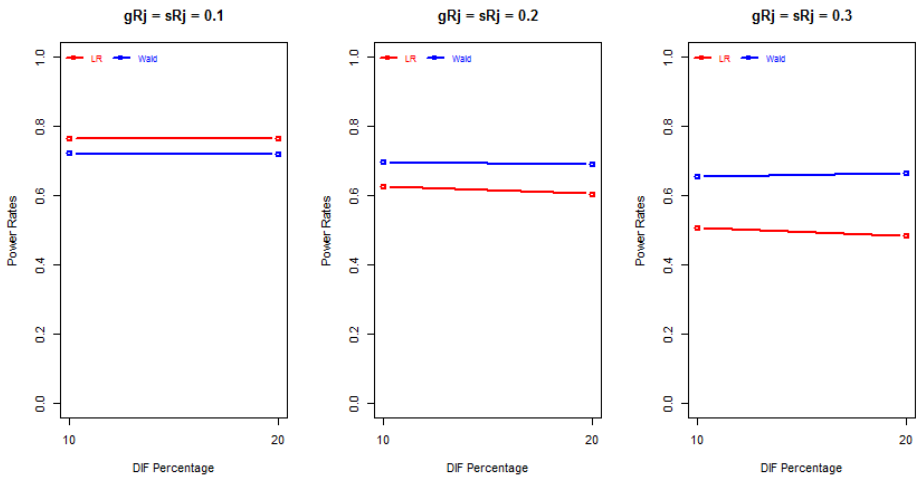
Graph 6: The Effect of the Number of Attributes on Power Rates of Methods

The results of the effects of the correlation between attributes on the power rates of methods were shown in Graph 7. According to Graph 7, the correlation levels between the attributes did not change the power rates of the Wald test. However, the power rates of the LR method increased with the increase in the correlation between the attributes. When the s and g parameter value was 0.1 and the correlation level between attributes were 0.2, the power rates of the Wald test method seem to be higher than the LR method. On the contrary, in cases where the correlation between the attributes were 0.5 and 0.8, the power rates of the LR method were higher than the power rates of the Wald test method. When s and g parameter values were 0.2 and 0.3, the power rates of the Wald test method were higher than the power rates of the LR method for all correlation levels between attributes. For all correlation levels between attributes, the power rates of LR method decreased with the increase of s and g parameter values.



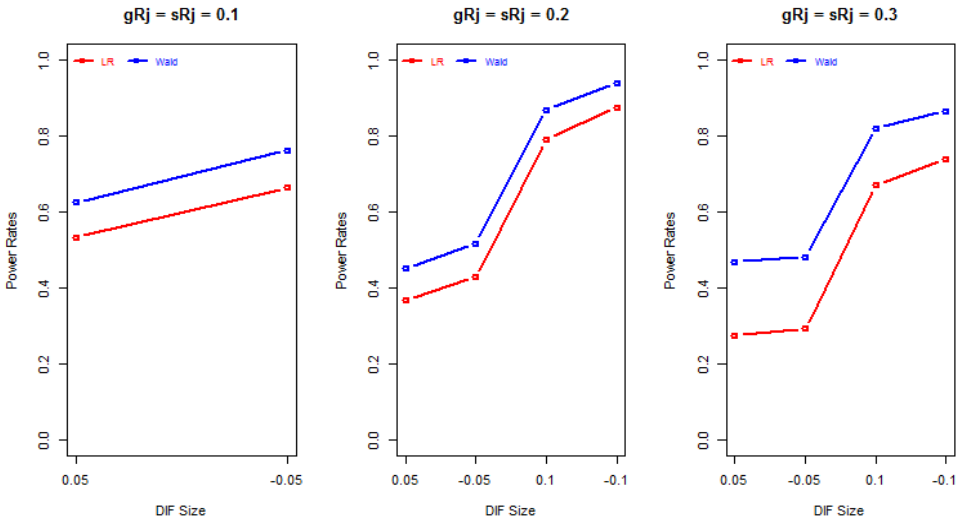
Graph 7: The Effect of the Correlation between Attributes on Power Rates of Methods

The results of the effects of the percentage of DIF items on the power rates of methods were shown in Graph 8. According to the Graph 8, when the s and g parameter value was 0.1, the power rates of the Wald test method were lower than the power rates of the LR method for both percentages of DIF items. When the s and g parameter values were 0.2 and 0.3, the power rates of the Wald test method were higher than the power rates of the LR method for both percentages of DIF items. For all percentage of DIF items, when the s and g parameter values increased, there was a decrease in the power rates of the LR method while there were slight differences in the power rates of the Wald test method.



Graph 8: The Effect of the Percentage of DIF Items on Power Rates of Methods

The results of the effects of the DIF sizes on the power rates of methods were shown in Graph 9. When the Graph 9 was investigated, it was observed that the power rates of the methods increased as DIF sizes increased for all the s and g parameters values. In addition to that, for all s and g parameters values and DIF sizes, the power rates of the Wald test method were higher than the LR method.



Grafik 9: The Effect of the DIF Size on Power Rates of Methods

CONCLUSION AND DISCUSSION

In this study, it was aimed to investigate the effects of various factors on the performance of the methods used in the determination of DIF in the DINA model. For this purpose, invariance of slip and guess parameters for focal and reference subgroups needed to be investigated. In order to determine DIF in the DINA model, several methods exist in the literature. Usability of these methods may vary according to several conditions and performance of these methods also need to be investigated accross these conditions.

In the DINA model, Logistic Regression and Wald test methods were the common methods which were used to determine the differential item functioning, and the Type I error and power rates of these methods in certain conditions needed to be investigated. In determining the Type I error rates of the methods, three sample sizes (500, 1000 and 2000), two number of attributes (4 and 5), three correlation between attributes levels (0.2, 0.5 and 0.8) and three reference group item parameters (0.1, 0.2 and 0.3) were manipulated. When the results obtained from the Type I error rates of the methods were investigated, it was observed that especially the s and g parameter values were effective factors on the Type I error rates of the methods. When the s and g parameter value was 0.3, it was observed that the Type I error rates of the Wald test method increased. It is consistent with the results of Hou et al. (2014). When the s and g parameter values were 0.1 and 0.2, it was observed that both the LR and the Wald test Type I error rates were close to each other. It was observed that the Type I error rates of the Logistic Regression method were not affected by the increase in sample size. However, when the s and g parameter value was 0.3, it was observed that the Type I error rates of Wald test method decreased dramatically with the increase in sample size and increased dramatically with the increase in the number of attributes. In addition to these, it was observed that the correlation between the attributes did not cause much change in the Type I error rates of the methods.

When the results of the power rates of the methods were investigated, it was observed that especially the sample sizes, DIF sizes and the s and g parameter values were effective factors. It is consistent with the results of Hou et al. (2014). The power rates of both methods increased with increasing sample size. It was observed that the increase in the percentage of DIF items did not change the power ratios of both methods. As the number of attributes increased, the power rates of the Wald

test method did not change much, but the power rates of the LR method decreased. While the level of correlation between attributes did not change the power rates of the Wald test method much, the power rates of the LR method increased with the increase in the correlation between the attributes.

In conclusion, it can be said that the Wald test method showed satisfactory results to detect DIF in many different conditions in the Cognitive Diagnostic Models. However, when the Wald test method is compared with the LR method, under some simulation conditions (i.e., when the s and g parameter values are high) the Wald test has inflated Type I error rates but in many conditions**,** it shows high power rates.

In this study, only the DINA model was used to simulate DIF conditions, differential item functioning can be investigated in further studies, by using other CDM models (DINO, GDINA etc.). In further studies, different DIF detection techniques can be used and the performances of these methods can be compared. In addition, further studies will be conducted with different factors or factor levels.

REFERENCES

Camilli, G. ve Shepard, L. A. (1994). *Methods for identifying biased test items*. London: Sage Publications.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34, 115-130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76, 179–199

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69, 333-353.

de la Torre, J., & Lee, Y. S. (2010). A note on the invariance of the DINA model parameters. *Journal of Educational Measurement*, *47*, 115–127.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. Nichol, S. Chipman, & R. Brennan (Eds.), *Cognitive diagnostic assessment* (pp. 361-389). Hillsdale, NJ: Lawrence Erlbaum.

Embretson, S. E. (1984). A general latent trait model for response processes. Psychometrika, 49, 175–186.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, 26, 333-352.

Hambleton, R K., Swaminathan, H. ve Rogers, H. J. (1991). *Fundamentals of item response theory*. London: Sage Publication

Hartz, S. (2002). Skills diagnosis: Theory and practice. User Manual for Arpeggio software. Princeton, NJ: ETS.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191.

Hou, L., de la Torre, J. D., and Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, 51, 98–125.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions and connections with nonparametric item response theory. *Applied Psychological Measurement,* 25, 258-272.

Li, F. (2008). *A modified higher-order DINA model for detecting differential item functioning and differential attribute functioning* (Doctoral dissertation). University of Georgia, Athens.

Li, X., and Wang, W. C. (2015). Assessment of differential item functioning under cognitive diagnosis models: the DINA model example. *Journal of Educational Measurement*, 52, 28–54.

Osterlind, S. (1983). *Test item bias*. Newbury Park: Sage Publications.

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

Rogers, S. J., & Swaminathan, H. (1993). A comparison of logistic regression and MH procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*, 105–116.

Tatsuoka, K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics*, 12, 55-73.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11(3)*, 287-305.

Templin, J. L., Henson, R. A., & Douglas, J. (2006). General theory and estimation of cognitive diagnosis models: Using Mplus to derive model estimates. Manuscript under review.

Zhang, W. (2006). *Detecting Differential Item Functioning Using the DINA Model*. Doctoral dissertations, University of North Carolina at Greensboro, Greensboro, NC.

Zumbo, D. B. (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 item scores*. Ottowa: Directorate of Human Resources Research and Evaluation, Department of National Defense.