KURAM VE UYGULAMADA EĞİTİM BİLİMLERİ EDUCATIONAL SCIENCES: THEORY & PRACTICE

Received: October 18, 2016
Revision received: December 19, 2016
Accepted: July 6, 2017
OnlineFirst: December 5, 2017

www.estp.com.tr
DOI 10.12738/estp.2017.6.0526 • December 2017 • 17(6) • 1977–2000

Research Article

Comparing Differential Item Functioning Based on Manifest Groups and Latent Classes*

Şeyma Uyar[1]
Mehmet Akif Ersoy University

Hülya Kelecioğlu[2]
Hacettepe University

Nuri Doğan[3]
Hacettepe University

Abstract

In this study, performance of differential item functioning (DIF) methods was compared under 36 different conditions based on latent classes and manifest groups. In the study, simulation conditions such as DIF-containing item rate, reference-focal group rate, DIF effect size and overlap ratio of manifest groups and latent classes were taken into consideration. To examine DIF, the Mantel–Haenszel (MH) method, which is a method related to the manifest group variable, was used within the framework of classical test theory and Lord's x^2 method and item response theory. Latent classes were determined using the model of multilevel mixture item response theory (MMIRT). Results show that data fit the MMIRT model with larger effect size and with a higher number of items containing DIF. When DIF effect size was 1.0, the power of MMIRT was found to be higher and the type I error rate was found to be lower in all overlap and DIF-containing item rates and reference-focal group conditions. While the rate of overlap was 90%, the power of MH and Lord's x^2 methods and type I errors were at acceptable levels under all conditions. It was observed that the power of MH and Lord's x^2 methods decreased as a result of a decrease in the overlap ratio for manifest groups and latent classes.

Keywords

Mixture distribution • Multilevel mixture item response theory • Latent class • Differential item functioning • Bias

* This article is an extension of corresponding author's doctoral thesis at Hacettepe University.

1 **Correspondence to:** Şeyma Uyar (PhD), Department of Educational Sciences, Faculty of Education, Mehmet Akif Ersoy University, Burdur Turkey. Email: syuksel@mehmetakif.edu.tr

2 Department of Educational Sciences, Hacettepe University, Ankara Turkey. Email: hulyaebb@hacettepe.edu.tr

3 Department of Educational Sciences, Hacettepe University, Ankara Turkey. Email: nuridogan2004@gmail.com

Citation: Uyar, Ş., Kelecioğlu, H., & Doğan, N. (2017). Comparing differential item functioning based on manifest groups and latent classes. *Educational Sciences: Theory & Practice, 17,* 1977–2000. http://dx.doi.org/10.12738/estp.2017.6.0526

A test item should be able to measure ability without involving characteristics of subgroups that consist of individuals. This is because individuals with equal abilities should be able to correctly answer an item at the same rate even though they are in different subgroups. If items included in the test provide more advantages for one group over another, the item is considered to be biased (Camili & Shepard, 1994; Mellor, 1995; Zumbo, 1999). Therefore, when developing a test, items should be examined in terms of item bias.

Item bias determination processes are carried out in two stages. The first stage is a statistical process during which item response distributions are examined in reference groups and focal groups established by considering observed variables (gender, country etc.) under equal ability levels (Cohen & Bolt, 2005; Steinberg & Thissen, 2006). In this distribution, differentiation in the probability of correct answers given provides the differential item functioning (DIF) of an item. Statistical properties of an item with DIF also vary among groups (Angoff, 1993; Clauser & Mazor, 1998). For this reason, experts should reveal in qualitative studies whether items with DIF are biased.

DIF analyses are usually carried out over reference and focal groups established based on manifest groups. In these studies, it is assumed that the characteristics of all participants are similar in the manifest group (De Ayala, Stapleton, & Dayton, 2002). In line with this assumption, an item with DIF is considered advantageous or disadvantageous for all individuals in a manifest group. Indeed, according to Samuelsen (2005), the reliability of results in DIF methods for the manifest group is affected by the assumption that a group consists of homogeneous communities. This is so in terms of its measured ability, and from the lack of consideration of the possibility that an item may contain DIF in the same group. This is because individuals in different subgroups (including gender, socioeconomic level, or culture) can be divided into latent classes that may be homogeneous with respect to ability (De Ayala et al., 2002; Samuelsen, 2005). A high level of overlap between these latent groups and manifest groups is low in probability. In other words, if members of a manifest group are also included in a single latent class, a 100% level of overlap can be considered in the manifest group and latent classes. This rate shows similarity between distinctive properties of the latent class and the manifest group. However, individuals of a group may also be members of another latent group. Accordingly, in these cases, especially when the ratio of overlap is less than 70%, obtained DIF results can be biased using only manifest group variables (Bilir, 2009; Samuelsen, 2005).

Hu and Dorans (1989) pointed out that girls tend to achieve lower scores than boys in the event of item removal. However, removal of an item also resulted in an increase in Asian American girls' scores compared with those of Latin and Asian American boys. Accordingly, though it seemed that girls' scores declined, scores in

Latin and Asian American latent class, within the subgroup of girls, increased. Cohen and Bolt (2002) pointed out that items showed DIF as per gender in their DIF method study on the manifest variable. However, latent class analysis results found out that about 50% of women and 40% of men were included in different classes. De Ayala et al. (2002), in a DIF study applied on classes determined through latent class analysis that three items showed DIF in the black race in the latent class but did not show DIF in the black race from the other class. These results suggest that the manifest variable method might determine an item as an item with DIF for all members even if this was not the case in reality.

Previous studies indicate that group homogeneity assumptions are not always met. Therefore, DIF method studies have emerged based on the latent class (Bilir, 2009; Cho, 2007; De Ayala, 2002; De Mars & Lau, 2013; Oliveri, Ercikan, & Zumbo, 2013; Samuelsen, 2005). Kelderman and Macready (1990) suggested that the latent class approach could be advantageous, and that the use of latent class variables enabled assessment without binding DIF to any variable sets.

The leading model for determining DIF by latent class is "Mixture IRT" (MIRT-Mixture Item Response Theory) within the scope of item response theory (IRT). This model classifies individuals into non-predictable latent classes based on their responses. Classes are homogeneous among themselves for their relevant characteristics and heterogeneous among classes. Furthermore, item difficulties are estimated differently for each latent class. This makes it easier to identify DIF (De Ayala et al., 2002). However, as in many models used in the field of statistics, MIRT also supposes that observations are independent of each other (Vermunt & Magidson, 2002). It is difficult to meet this assumption, especially for applications in the field of education. Therefore, different model types are suggested for hierarchical data in which students are clustered in classes, classes are clustered in schools and schools are clustered in cities (Stevens, 2009; Aspourov & Muthen, 2008). Cho and Cohen (2010) suggested advocated use of the "Multidimensional MIRT" model (MMIRT) since this model provided more precise information about group membership based on individuals' response patterns. The model was more effective in determining DIF by including effects of student and school level on the model. Nevertheless, performance of methods in determining DIF can be sensitive to a variety of factors or interactions between these factors.

Many studies reveal that DIF methods are affected by various variables such as test length, sample size, ratio of items with DIF and effect size of DIF (Clauser, Mazor, & Hambleton, 1993; Cho, 2007; Samuelsen, 2005). In these studies, performance of methods can be examined through comparison in terms of type I errors and statistical power based on various conditions (Finch, 2005; Kim, 2010; Naranayan

& Swaminathan, 1994). Generally, it was a common finding in studies that IRTDIF-based methods were more effective. However, it was not easy to meet assumptions required by the model (Narayanan & Swaminathan, 1996). One method frequently used in DIF studies is the Mantel–Haenszel (MH) method based on classical test theory (CTT). It was concluded in these studies that type I errors of the MH method were low in small and large samples and in cases when ability distributions did not vary. Furthermore, this method provided acceptable results under many conditions (Roussos & Stout, 1996; Prieto, Barbero, & Luis, 1997; Shealy & Stout 1993). With regard to the literature, it was suitable to compare of DIF in terms of latent class and manifest group variables with powerful IRTDIF-based methods and CTTDIF-based methods, which meet assumptions more easily.

Differential Item Functioning

Differential item functioning refers to undifferentiation of item characteristic curves and possibilities of correct answers for items in groups in situations when ability is examined (Clauser & Mazor, 1988; Li & Stout, 1996; Naranayan & Swaminathan, 1996). Item characteristic curves graphically demonstrate the possibilities related to responses that an individual with a specific level of ability may provide (Hambleton & Swaminathan, 1985).

DIF occurs in two ways: uniformly and non-uniformly (Mellenberg, 1982). Uniform DIF emerges in situations when item characteristic curves are parallel, and provides benefits to only one of the groups for each level of ability. Item discriminations do not vary among groups. Non-uniform DIF is defined as functioning of an item in favor of one group at some ability level and in favor of the other group in other ability levels throughout the ability scale. Discrimination and difficulty parameters for items differ for the reference group and focal group (De Ayala et al., 2002; Zumbo, 1999). There are many different methods used to determine DIF. In this study, MH, one of the methods based on CTT, and Lord's x^2, one of the methods based on IRT, are explained in detail.

Mantel–Haenszel Method

Mantel–Haenszel is one method used to determine uniform DIF. This method is attained from scores from dichotomously scored items by two groups with the same level of ability (the focal and reference groups). Determination depends on the difference between "odd" values calculated through dividing the possibility of realization for an event by the possibility of its non-realization (Mertler & Vannatta, 2005). Ability groups are established with individuals who achieve similar scores based on total test scores. For each ability level, a 2 x 2 cross table is created (Holland & Thayer, 1988). In Table 1.1, the number of individuals with correct and wrong answers for an item in the reference group and the focal group and the total number of respondents are given.

Table 1
Data Layout according to MH Technique

Group	Correct	False	Total
Reference	A_j	B_j	n_{Rj}
Focal	C_j	D_j	n_{0j}
Total	m_{1j}	m_{0j}	T_j

The likelihood ratio (is calculated with the help of values in Table 1. It has a value lower than 1, then the item is seen to offer an advantage to the focal group. If, however, it is larger than 1, the possibility of the reference group giving a correct answer is higher. Camilli and Shepard (1994) suggested that (delta) statistics could be used by taking −2.35 times the natural logarithm of to facilitate interpretations. The resulting statistics are interpreted as the size of DIF effect that determines the level of DIF. Dorans and Holland (1993) pointed out that if is true for effect sizes, the item contains negligible DIF at A level or does not contain DIF at all. However, if , the item contains B level or middle level DIF, and if , the item contains C level, or a high level of DIF.

Lord's x^2 (Chi-Square) Method

To determine uniform and non-uniform DIF, Lord (1980) suggested using the x^2 method based on a suitable item response model (Maij de Meij, Kelderman, & van der Flier, 2010; Wiberg, 2007). This method is based on comparison of item parameters among groups. The x^2 statistic is calculated with the help of the difference between calculated item parameters and a variance-covariance matrix related to this difference (Camilli & Shepard, 1994). The obtained x^2 statistic adheres to the Chi-square distribution with "1" degree of freedom. It is concluded that when the x^2 statistical value exceeds the critical value, the item contains DIF based on the relevant level of significance.

Latent Variable Modeling Approaches

In the social sciences, it is known that many properties cannot be observed or measured directly. These properties, also called latent variables, can be explained indirectly based on statistical models that associate observed variables with latent variables (Jöreskog & Sörbom, 1993; Skrondal & Hesketh, 2004). Types of analysis based on latent models are divided into different categories based on the continuous or discrete structure of observed and latent variables. Item response theory, latent class analysis, factor analysis, and latent profile analysis are among the models that are used to define latent variables. To obtain reliable results, it is important to determine the model to be used based on scale and variable type. Models have different features and fields of application, and can be used together to reveal certain characteristics (Cho, 2007).

Item Response Theory

According to IRT, an individual's ability related to any analyzed property can be estimated based on his/her responses to an item. The relationship between an individual's observed test performance and latent property for this performance is defined as the item response model (Hambleton & Swaminathan, 1985). In IRT, an individual's score from the test can be determined using mathematical models. Within the framework of IRT, models are divided into normal and logistic, but logistic models are mainly preferred. These models are called two-parameter, three-parameter, and one-parameter models.

In the three-parameter logistic model, there are difficulty, discrimination (slope) and guessing (c) parameters. Lord (1968; 1980) defines the c parameter as the likelihood that an individual will correctly answer a question based on their lowest level of ability rather than defining this parameter as a guessing parameter. The two-parameter logistic model is based on the assumption that chances of success are zero. In addition to item difficulty, the item discrimination parameter is also included in the model. Using these two parameters, individuals' abilities are estimated (Hambleton, Swaminathan, & Rogers, 1991). A one-parameter model is a special form of the two- and three-parameter logistic models (Hambleton & Swaminatthan, 1985). This model has a form of assumption such that success chance is zero and each item has the same level of discrimination power (an average a parameter value estimated for all items). In this model, only the item difficulty parameter is considered, and ability is estimated based on this parameter (Hambleton et al., 1991). A special form of one-parameter logistic model is the Rasch model. Here discrimination parameter is considered to be equal and 1 for all items.

Mixture Distribution Models

Mixture distribution models are used to model groups with a heterogeneous structure with two or more components based on related properties (McLachlan & Peel, 2000). There is indication that multiple-mode data is more successful in modeling compared to classical statistical models. Initial studies conducted toward a mixture model showed that adaptation of each subgroup separately into a normal distribution model included fewer errors when compared against modeling heterogeneous data with normal distribution because the related latent property might have a different distribution in the subgroups (Çalış, 2005; Frühwirth-Schnatter, 2006; Sarı, 2012). In Figure 1, one-component normal distribution and two-component normal distribution curves are shown with regard to crab-type data that was analyzed by Pearson (1984) to examine the mixture model.

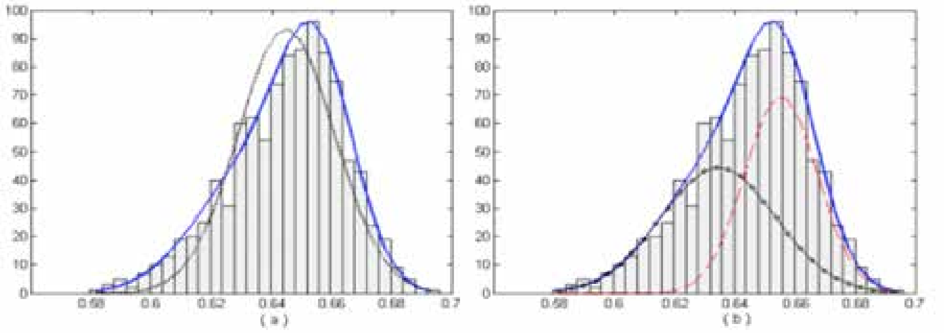

Figure 1. a) One-Component Normal Distribution Model (dashed line) and Two-Component Mixture Normal Distribution Model (continuous line) b) Normal Distribution Curves for Each Component and Two-Component Normal Distribution Curve (McLachlan & Peel, 2000).

According to Figure 1.b, it is estimated that there may be two separate crab types for young crab data with an asymmetric structure. Therefore, Pearson (1984) suggests that using the normal distribution seen in the discrete line of Figure 1 is not correct. That is, it might be more appropriate to define this data with a mixture of models using two separate components (Çalış, 2011; McLachlan & Peel, 2000).

An observed variable in a mixture distribution model is expressed using conditional probability functions. When the observed variable is discrete (countable), this causes the mixture model to be called the finite mixture model. In addition, there are uncountable mixture models. Latent class analysis that helps to separate heterogeneous groups is a type of analysis based on finite mixture distribution models (Vermunt & Magidson, 2005).

Latent Class Analysis

Latent class analysis (LCA) was first used by Lazarsfeld (1950) to explain the heterogeneity of a group in a study that included response patterns consisting of two-category items. Using LCA, for each of V number of observed variables, classes they belong to and number of classes (T) suitable for data set are determined. To determine the number of latent classes, the model should first be determined. Model selection starts with determining the number of classes. The process of deciding on the number of classes requires hypothesis testing. A hypothesis is established from simple to complex models. First, the process starts with the basic model (zero hypothesis) suitable for the $T = 1$ class that expresses mutual independence among variables. Zero hypothesis expresses the situation in which the model fits with the data. Accordingly, there are not any relationships between observed variables, and LCA is not necessary. Rejection of the zero hypothesis refers to differentiation of parameters into some subgroups (Vermunt, 2005). In this case, the latent class model with *T = 2* class is tested. At each turn, another dimension is added by increasing the number of latent classes. This process continues until the simplest model is obtained, namely the model with the fewest parameters.

Mixture Item Response Theory

Using IRT and LCA together yields "mixture item response theory" (MIRT) (Cohen & Bolt, 2005). A mixture model is defined by Rost (1990) as "Mixture Rasch" model. This is a combination of the latent class and Rasch model that can separate the number of latent classes assumed to be infinite in the universe according to individual response patterns. In the Mixture Rasch model, latent classes are established by considering observed variables in a multivariate structure. Item parameters are simultaneously estimated in accordance with individual ability and the class he/she belongs to (Alexeev, Templin, & Cohen, 2011; Cohen & Bolt, 2005; Mislevy & Verhelst, 1990; Rost, 1990, 1997). In the Mixture Rasch model, it is admitted that each latent class fits the Rasch model, but that classes have different item difficulty parameters. Parameters estimated with this model are specific to latent class. According to this model, the formula related to the possibility of a correct answer is as follows.

$$P(y_{ijg} = 1|g, \theta_{jg}) = \frac{1}{1 + exp[-(\theta_{jg} - \beta_{ig})]} \qquad (1)$$

In Equation 1, refers to the index of a specified latent class; j = 1, ...,*J* refers to the index of specified responders; θ_{jg}: *j.* refers to the individual's latent ability in the latent class; and *β* refers to the difficulty parameter of item in class g. In the Mixture Rasch model, the structure of ability is given as follows.

$$\theta_{jg} \sim N(\mu_g, \sigma_g^2) \qquad (2)$$

According to Equation 7, ability has a normal distribution with and parameters. μ indicates ability average with class features and σ indicates ability variance with class features. According to Rost (1990), the greatest advantage of using the model is its ability to concurrently calculate individuals' abilities for the same item and also to reveal the differences between individuals by separating them into their latent classes based on their response patterns (Cho & Cohen, 2010). At the same time, the Mixture Rasch model can be developed according to the 2-PL and 3-PL model (Bolt & Cohen, 2005; Finch & Finch, 2013).

Multilevel Models and Multilevel Mixture Item Response Theory

Multilevel models (Hierarchical Linear Models-HLM) allow formal application of multilevel data structures frequently used in education and psychology (Bryk &

Raudenbush, 1992; Longford, 1993). In this way, researchers are able to observe the effects of different variables such as school and curriculum at lower levels (e.g., students). When multilevel structure is ignored, non-applicable results emerge in item parameters, standard errors and DIF estimates unless the variance from a variable is zero or near-zero at a certain level. Therefore, in DIF studies, multilevel methods are applied (Finch & Finch, 2012).

HLM allows more precise prediction of standard errors for model item parameters which emerge when HML is combined with item response theory (Fox, 2005; Maier, 2001, 2002). Kamata (2001) developed a three-level IRT model (Hierarchical General Linear Models-HGLM) for dichotomously scored items. In the model, the first level refers to item, the second level refers to student and the third level refers to school levels.

The first level of HGLM represents the measurement model. At this stage, regression coefficients are determined for all items. The second level of the model refers to inclusion of student level. At the third level, school variable is added to the model. Ability estimation is carried out based on school level.

Defining the hierarchical structure of the data set in the model offers a great advantage in terms of searching for the underlying factor of DIF. Item parameters are simultaneously estimated with an individual's ability and the class he/she belongs to (Alexeev et al., 2011; Cohen & Bolt, 2005; Mislevy & Verhelst, 1990; Rost, 1990, 1997). However, it is interpreted as a disadvantage that HGLM does not provide information about group membership of individuals together with determination of DIF (Cho, 2007). Therefore, Cho (2007) defined Multilevel Mixture Rasch model (MMIRT) to analyze DIF.

Unlike HGLM, it is possible to create latent classes at student and school level by means of MMIRT. DIF comparisons can also be made among these latent classes. MMIRT is expressed with the following equation when individuals are grouped in schools (level 1) or classes (level 2).

$$P\left(y_{ijtg}=1 \mid g,k,\theta_{jtgk}\right)=\frac{1}{1+\exp\left[-\left(jgk-\beta_{igk}\right)\right]} \tag{3}$$

g represents the first level latent class (student); *k* represents the second level latent class (school); θ_{jtgk}, *g* and *k* represent *j* individual's latent ability from latent classes and β_{igk}, school; and shows the difficulty level of *j*. item for *g* and *k* latent classes.

Cho (2007) indicates that MMIRT can be used in three different situations in DIF applications.

Special case 1: Occurs when item and ability parameters can be estimated separately for student and school levels. When item parameters vary in classes that are established based on student and school levels, parameters can be estimated for latent classes at all levels.

Special case 2: Occurs when item and ability parameters do not vary for different school level classes. This model is functional since it includes different properties of students in the multilevel model. In this model, item difficulty values only vary as per classes at the student level.

$$P_B(y_{ijtgk} = 1 \mid g, k, \theta_{jtg}, \theta_{jt}) = \frac{1}{1+\exp\left[-(\theta_{jg}+\theta_{jt}-\beta_{ig})\right]} \tag{4}$$

In Equation 4, it is apparent that the model does not include a k index meaning that equal estimates will be applied for school level latent classes. θ_{jt} shows the ability of an individual who is in the g latent class. When this model is used, similar estimates can be achieved for the school level latent classes. When considered in this respect, MMIRT provides information about DIF at the school level. Differentiation among students emerges only with a multilevel data structure (Asparouhov & Muthen, 2007; Cho, 2007).

Special case 3: Occurs when item and ability parameters do not vary among student level classes. This case was suggested by Vermunt (2007) to determine DIF in school level latent classes.

The Purpose and Importance of This Study

In many studies, it has been observed that DIF studies can be carried out over manifest groups. However, it is impossible for individuals in these groups to resemble each other completely. Therefore, results may be inadequate when obtained from DIF studies carried out based on only manifest groups. If the manifest group and the latent variable do not coincide completely, especially if this ratio is lower than 70%, one might suggest that DIF studies be carried out over latent classes with more objective estimates obtained to determine the real reason. Besides, since data has a multilevel structure, it requires a multilevel DIF approach. On the other hand, awareness of conditions effecting DIF methods and knowing when more precise results are obtained facilitates the determination of items with DIF and enables achievement of valid and reliable tests. Especially, since biased items in large-scale examinations may affect decisions based on test scores, it is important that researchers use the most accurate method for various conditions. Therefore, selection of DIF methods by determining a model suitable to data structure will provide experts with the opportunity to examine more items in terms of bias. In accordance with the objective of this research in which related issues and suggestions are considered, an answer to the following problem statement is sought.

Do the statistical power and type I errors related to DIF determined based on the manifest group and DIF determined based on the latent class differ when DIF effect size, overlap ratio, DIF-containing item ratio and reference-focal group rate change?

Method

Data

In this study, simulation data was used to analyze the performance of DIF techniques. Data used in the study had a multilevel structure. Data was acquired by selecting 50 students from 100 different school samples by considering cases in which students from different classes took an exam in large-scale tests. In the study, there was a two-category manifest group variable (gender etc.), a two-category student level variable (economic level etc.) and a two-category school level variable (school location etc.).

It was observed that researchers worked with different numbers of items in studies in which DIF analyses were carried out through the MIRT model. Cho, Cohen, and Kim (2006) suggested working with more than 10 items for the MIRT model (as cited in Bilir, 2009). This study was conducted over 20 items. Cho (2007) indicated that the minimum sample size for MMIRT model could be 1000. Samuelsen (2005) indicated that the power of MIRT in DIF was insufficient when sample size was lower than 2000. Cho (2007) and Zhu (2013) carried out their studies over large samples of their work (8000 and 6000 items). For comparisons with previous surveys and more powerful estimates using MMIRT, in this study, sample size was determined as 5000. N~(0,1) unit was established in which the average item difficulty was zero. Standard deviation was set at one and the abilities of the manifest group and latent classes were zero. Uniform DIF was analyzed by considering the case in which only item difficulty parameters varied among latent classes.

Simulation Conditions

Percentage of DIF-containing items. In this study, a 20% and 40% DIF item ratio was used. Four items (items 3, 4, 10, and 16) were produced when 20% of items in the test contained DIF, and eight items were produced (items 3, 4, 10. 16, 17, 18, 19, and 20) when 40% of items contained DIF.

DIF effect size. For the MIRT model, Samuelsen (2005) worked with 0.2 (low), 0.8 (medium), and 1.2 (high) effect sizes. Cho and Cohen (2010), working with MMIRT, carried out simulations with effect sizes of 0.4, 0.6, 0.8, 1.0 and 1.2. Bilir (2009) studied the negative and positive values of 0.5 and 0.7 effect sizes for the Mixture Rasch-MIMIC model. In this study, 0.5, 0.7, and 1.0 effect sizes were used with difficulty parameters set at low, middle, and high levels.

Overlap ratio of manifest group and latent classes. Samuelsen (2005) stated that powerful DIF methods were required for cases when overlap ratio was lower than 70%. Therefore, 90%, 70%, and 50% ratios were used by considering high, middle, and low levels of overlap rate with latent classes of a manifest group variable.

Reference and focal group ratio. In this study, the simulation condition was determined as a 50:50 rate in which sample size was equal to reference and focal groups, and an 80:20 ratio in which sample size was not equal. The number of individuals according to the overlap ratio is summarized in Table 2.

Table 2
Number of Individuals in Manifest Group and Latent Classes as per Overlap Rate

		Reference and focal group rate			
		50:50		80:20	
Overlap Rate		Latent Class 1	Latent Class 2	Latent Class 1	Latent Class 2
90%	Group 1	2250	250	3600	400
	Group 2	250	2250	100	900
%70	Group 1	1750	750	3200	800
	Group 2	750	1750	300	700
%50	Group 1	1250	1250	2000	2000
	Group 2	1250	1250	500	500

Data Generation

Based on related conditions, the data set was produced by code written in C# (C sharp) software based on the 1-PL model. Phases of data generation are described below in turn.

(i.) First of all, parameters of ability and item difficulty used in data generation were produced in 3 range to show standard normal distributions. With the aim of creating answers for the first student level, difficulty parameters (beta) values for 20 items were used. Answer patterns for the second student level latent class was established by adding effect sizes to difficulty parameters of items including DIF. Changing difficulty and ability parameters in latent classes at the school level was disallowed. Consequently, similar estimates could be achieved for school level latent classes. *(ii.)* For each item, U[0.1] random number was produced that gave a uniform distribution (taking a value between 0 and 1). *(iii)* A student's possibility of giving correct answers to the item was calculated based on the special case 2 formula of MMIRT. *(iv)* Obtained correct answer rates were compared with random numbers that were obtained at the third step. At this stage, when the random number was lower than the correct answer rate, the final value of the item was determined as 1 by considering the fact that the individual answered the item correctly. When the random number was greater than the correct answer rate, a zero value was assigned to the item. This value indicated that the individual had given th wrong answer to that item. In Table 3, item difficulty values used during data generation are given.

Table 3
Item Difficulty Values Used in Data Generation

	C1	C2				C1	C2		
Item		DIF Effect Size					DIF Effect Size		
		0.5	0.7	1.0		Item	0.5	0.7	1.0
1	0.3				11	0.47			
2	−0.8				12	1.77			
3	−0.6	**−0.1**	**0.1**	**0.4**	13	−0.47			
4	1.4	**1.9**	**2.1**	**2.4**	14	−0.9			
5	−2.0				15	0			
6	0				16	0.63	**1.13**	**1.33**	**1.63**
7	−0.01				17	−1.56	**−1.06**	**−0.86**	**−0.56**
8	0.15				18	0.7	**1.2**	**1.4**	**1.7**
9	−1.16				19	0.97	**1.47**	**1.67**	**1.97**
10	0.1	**0.6**	**0.8**	**1.1**	20	−0.07	**0.43**	**0.63**	**0.93**

C: student level shows difficulty values only for DIF-containing items for C2.

When the rate of items with DIF was 20%, items 3, 4, 10, and 16 were produced to indicate DIF. When the rate of items with DIF was 40%, items 3, 4, 10, 16, 17, 18, 19, and 20 were produced to indicate DIF. Difficulties for other items did not vary.

Data Analysis

Analyses of MMIRT were carried out in accordance with the 1-PL model and special case 2. In selection of the model appropriate to the data, it is possible to use Akaike's information coefficient (AIC) and Bayesian information criterion (BIC) statistics can be used. A BIC value is preferred in comparison of 2- and 3-parameter MIRT models, namely, in the selection of simple models. An AIC value is more useful in selection of complex models (Cho, 2007; Zhu, 2013). The most appropriate model for data is selected in the case when the AIC value is the lowest. In this study, regardless of the overlap rate, lower AIC values were obtained with the 1PL model while the effect size was 0.5 and the rate of items with DIF was 20% (AIC 1pl model: 115435.537, AIC MMIRT: 115438.619). In cases where the rate of items with DIF was 40% and the effect size was 1, data showed a better fit with the MMIRT model (AIC 1PL model: 113001.762, MMIRT: 112987.432). In analyses for the MMIRT model, Mplus software was used through R software (Hallquist, 2015). As a result of the analysis, two difficulty values were calculated for each item. To analyze DIF among student level latent classes, estimated item parameters with class properties were considered as a reference group for one class and as a focal group for the other class. These were acquired in a manner such that the first student level would be the first school level (1-1) and the second student level would be the second school level (2-1).

It is possible to use different methods to compare parameters among groups in terms of DIF that are estimated via the latent class property. These methods include differences in difficulty parameters, the marked and unmarked field index, Lord's x^2 and MH techniques. In this study, coefficient suggested by Roussus, Scnipke, and Pashley (1999) was used:

$$\Delta = -2.35\ln(\alpha) = -2.35\ln\left[e^{-1.7a(b_R-b_F)}\right] = 4\alpha((b_R - b_F)$$

To analyze DIF, MH method from the CTT and the Lord's x^2 technique from the MIRT were used to interpret the manifest group variable. Analyses relating to this method were carried out in "*difR*" library in R 3.1.2 software. The Lord's x^2 and the MH x^2 values acquired as a result of analysis were analyzed, and items found to be significant as per a 0.5 level of significance were assessed as DIF-containing items. To determine DIF level, and statistics were used. Based on the size of this value, it was determined whether the item showed DIF at the A, B, or C level. To increase the reliability of analysis results, DIF comparison was repeated for different data sets. The number of repetition required was limited to 50 times. In repeated analyses, variance analysis (ANOVA) was used to determine whether acquired type I errors and power values varied with method used. Using factorial ANOVA, the effect of conditions on error and statistical power was tested. DIF performance was determined according to the manifest group, and the latent class was interpreted within the scope of suitability of data to the model.

Findings

Type I errors and power rates were acquired with DIF determined based on the manifest group, while the latent class was given in research questions according to overlap amount.

Table 4

Type I Error and Power Rates for Cases Where Overlap Rate is 90%

			Type I Error			Power (%)		
ES	R/F	DIR	MH	LORD	MMIRT	MH	LORD	MMIRT
0.5	50:50	4 (20%)	0.033	0.033	0.495	100.0	100.0	75.00
		8 (40%)	0.043	0.035	0.250	100.0	100.0	89.50
	80:20	4 (20%)	0.039	0.034	0.510	89.50	89.50	71.50
		8 (40%)	0.072	0.082	0.270	84.50	83.00	72.75
0.7	50:50	4 (20%)	0.055	0.044	0.128	100.0	100.0	100.0
		8 (40%)	0.038	0.043	0.120	100.0	100.0	100.0
	80:20	4 (20%)	0.046	0.038	0.253	100.0	100.0	91.00
		8 (40%)	0.067	0.062	0.140	100.0	100.0	100.0
1	50:50	4 (20%)	0.046	0.039	0.064	100.0	100.0	100.0
		8 (40%)	0.130	0.082	0.033	100.0	100.0	100.0
	80:20	4 (20%)	0.031	0.022	0.070	100.0	100.0	100.0
		8 (40%)	0.028	0.033	0.053	100.0	100.0	100.0

R/F: Reference-Focal Group Rate, ES: Effect size, DIR: DIF-Containing Item Rate

Findings of Research Question 1

How do type I error and power rates of DIF methods vary when overlap ratio between manifest group and latent classes is 90%?

Type I error and power rates are given in Table 4 as calculated through MH, Lord's x^2 and MMIRT, with a 90% overlap existing between the manifest groups and latent classes depending on the first research question.

According to Table 4, the type I error rate of the MH method was within Bradley's flexible reference range ($0.025 \leq$ type I error rate ≤ 0.075) in all conditions with 0.5 and 0.7 effect sizes. It is only when the DIF effect size is increased to 1 that the reference-focal group rate is not distributed equally and the DIF item rate is 40%, it shows a higher rate of error (0.13). The Lord's x^2 method tends to give smaller error values that are generally similar to the MH method. The maximum error value (0.82) was found in the 40%-DIF item rate condition when the effect size was 0.5 with the unequal reference-focal group rate, and in 40% DIF item rate condition when the effect size is 1 with equal reference-focal group rate. While the DIF with estimated parameters based on MMIRT showed a higher error at the 0.5 and 0.7 effect sizes, the lowest error (0.033) was shown at the equal group rate when the DIF item rate was 40%.

Compared in terms of powers in the determination of DIF, MH (84.5), and Lord's x^2 (83) methods were shown to have the lowest power in the condition in which the effect size was 0.5 with the unequal group rate and the 40% DIF item rate. The power of these methods in all other conditions was at high levels. On the other hand, the MMIRT showed the lowest power under the condition of the 0.5 DIF effect size with the unequal group rate and the 20% DIF item rate. When the DIF effect size was 0.7, DIF power determined with MMIRT was found to be 90% at the unequal group rate and 20% at the DIF item rate. Other than this, it was found to be 100% in all 0.7 and 1.0 effect sizes.

Findings of Research Question 2

How do type I errors and power rates of DIF methods vary when the overlap ratio between the manifest group and the latent classes is 70%?

Analyzing Table 5 shows that the highest type I error (0.133) using the MH method was acquired under the condition of the 0.7 effect size, unequal group rate and the 40% DIF item rate. The MH method showed the lowest error (0.02) when the DIF effect size was 1.0, group rates were equal and the DIF item rate was 40%. Lord's x^2 method, similarly to MH method indicates the highest error (0.137) when effect size is 0.7, distribution of groups is not equal and DIF item rate is 40%. The lowest error value (0.03) obtained via this method was achieved when the effect size was 1.0 and group distributions wer equal. It was observed that the highest error value (0.50) for

Table 5
Type I Error and Power Rates for Cases Where Overlap Rate is 70%

			Type I Error			Power (%)		
ES	R/F	DIR	MH	LORD	MMIRT	MH	LORD	MMIRT
0.5	50:50	4 (20%)	0.069	0.059	0.421	81.50	76.00	77.00
		8 (40%)	0.105	0.105	0.292	53.00	51.00	84.70
	80:20	4 (20%)	0.08	0.067	0.507	44.50	41.00	53.50
		8 (40%)	0.092	0.088	0.318	28.75	25.50	88.25
0.7	50:50	4 (20%)	0.050	0.041	0.120	97.50	97.00	96.50
		8 (40%)	0.078	0.077	0.055	92.30	90.50	100.0
	80:20	4 (20%)	0.055	0.050	0.250	64.00	58.50	91.50
		8 (40%)	0.133	0.137	0.103	44.30	37.30	100.0
1	50:50	4 (20%)	0.034	0.030	0.063	100.0	100.0	100.0
		8 (40%)	0.020	0.030	0.032	100.0	100.0	100.0
	80:20	4 (20%)	0.040	0.035	0.061	89.00	87.50	100.0
		8 (40%)	0.105	0.123	0.047	72.75	66.50	100.0

MMIRT was obtained when the effect size was 0.5, the reference-focal group rate was not equal and the DIF item rate was 20%. Error related to DIF as per parameters that are estimated according to MMIRT obtained the lowest value (0.032) when the DIF effect size was 1.0, group rates were equal and the DIF item rate was 40%.

When compared in terms of power, and after the overlap rate fell to 70%, it was found that the MH method (28.75%) and Lord's x^2 (25.5%) showed the lowest power in situations where the DIF effect size was 0.5, group rates were 80:20 and the DIF-containing item rate was 40%. These methods have 100% power only when the effect size is 1.0 and group distribution is equal. The power of MMIRT indicates the lowest value (53.5%) under the condition of 0.5 effect size, unequal group rate and 20% DIF item rate. When effect size reaches 1, power values obtained with MMIRT reached 100% in all circumstances.

Findings of the Research Question 3

How do type I errors and power rates of DIF methods vary when the overlap ratio between manifest group and latent classes is 50%?

According to Table 6, when the overlap rate falls to 50%, error values related to MH and Lord's x^2 methods remain within acceptable limits. The MH method gives the lowest error (0.023) when effect size is 1.0, the group rate is not equal and the DIF item rate is 40%. Lord's x^2 method, on the other hand, exhibits the lowest error value in cases where effect size is 0.7, group rates are equal and the DIF-containing item rate is 40%. For MMIRT, it can be said that the lowest error (0.017) is obtained when the effect size is 1.0, the group rate is equal and the DIF-containing item rate is 40%.

The power of DIF determined via MH and Lord's x^2 methods was found to be very low and close to zero. MMIRT had the lowest power (60%) when the effect size

Table 6
Type I Error and Power Rates for Cases Where Overlap Rate is 50%

			Type I Error			Power (%)		
ES	R/F	DIR	MH	LORD	MMIRT	MH	LORD	MMIRT
0.5	50:50	4 (20%)	0.063	0.043	0.358	6.25	3.75	60.00
		8 (40%)	0.043	0.028	0.208	7.00	5.25	92.30
	80:20	4 (20%)	0.058	0.051	0.410	4.50	4.50	75.00
		8 (40%)	0.035	0.033	0.270	2.50	2.25	91.25
0.7	50:50	4 (20%)	0.051	0.038	0.298	5.00	7.50	89.50
		8 (40%)	0.03	0.018	0.070	5.30	2.75	100.0
	80:20	4 (20%)	0.041	0.036	0.220	2.00	2.00	92.50
		8 (40%)	0.056	0.050	0.130	2.50	1.75	100.0
1	50:50	4 (20%)	0.036	0.021	0.051	3.50	3.00	100.0
		8 (40%)	0.065	0.033	0.017	3.75	2.50	100.0
	80:20	4 (20%)	0.038	0.030	0.073	3.00	3.00	100.0
		8 (40%)	0.023	0.023	0.056	4.80	4.50	100.0

was 0.5, group rates were equal and the DIF item rate was 20%. When the effect size reached 1, DIF determined with estimated parameters based ın latent classes were found at the 100% level.

The results of ANOVA applied to determine whether there were any statistically significant differences between averages of type I errors and power rates calculated for the manifest variable and latent class methods are given in Table 7.

Table 7
Differentiation Status of Type I Error and Power Rates According to Methods

		Sum Of Squares	sd	F	Significant Difference
Type I Error	Between-Group	.427	2	28,193	MH-MMIRT; Lord x^2 -MMIRT
	Within-group	.795	105		
Power	Between-Group	27795.873	2	11.004	
	Within-group	132617.7	105		

$^{**}p < .001$.

Analyzing Table 7, it can be said that type I error averages of DIF methods significantly varied ($F_{2,105,0.001}$ = 28.193). According to multi-comparison test results, it was found that the type I error rate of MMIRT method (.187) was significantly higher than the MH method (.057) and Lord’s $x^{2 \text{ method (.049).}}$ There were no significant difference between type I error value magnitudes of MH and Lord’s x^2 methods.

According to ANOVA results, there was a statistically significant difference between power value averages of methods $F_{2,105,0.001}$=28.193). The power of MMIRT (91.468) was found to be significantly higher than MH (58.117) and Lord’s x^2 (56.794) methods. The power of methods based on manifest group were not statistically significant.

Factorial ANOVA was applied to determine the effect of variables on type I error and power rates, and results are given in Table 8.

Table 8

Analysis of type I error and power rates according to methods and variables

			MH		LORD		MMIRT	
		sd	*F*	η^2	*F*	η^2	*F*	η^2
Type I error	Reference-Focal Group Rate (RF)	1	0.244	0.008	2.209	0.071	4.344**	0.130
	Effect Size (ES)	2	0.416	0.028	1.243	0.079	101.048**	0.875
	DIF-Containing Item Rate (DIR)	1	4.477**	0.134	7.962**	0.215	35.810**	0.553
	Overlap Rate (OR)	2	3.143**	0.05	8.649**	0.374	0.221	0.015
	OR*ES	2	.606	0.70	0.339	0.041	0.592	0.069
	OR*DIR	1	.367	0.022	1.748	0.098	0.045	0.003
	RO*OR	2	1.561	0.163	1.356	0.145	0.625	0.072
	ES*DIR	2	.484	0.057	0.223	0.027	11.086**	0.581
	ES*OR	4	1.089	0.214	0.362	0.083	1.948	0.328
	DIR*OR	2	1.518	0.16	4.686**	0.369	0.051	0.006
	OR*OR*ES	6	1.588	0.614	1.450	0.332	1.026	0.506
	OR*ES*DIR	3	.709	0.262	0.790	0.283	0.146	0.068
	OR*ES*DIR	6	.910	0.477	1.569	0.611	0.653	0.395
			MH		LORD		MMIRT	
		sd	*F*	η^2	*F*	η^2	*F*	η^2
Power	Reference-Focal Group Rate (RF)	1	1.46	.048	9,368	.24	.938	.031
	Effect Size (ES)	2	4,406**	.223	4.3**	.229	40.157**	.735
	DMF-Containing Item Rate (DIR)	1	.42	.014	1.754	.057	13.75**	.32
	Overlap Rate (OR)	2	41.825**	.743	169.6**	.92	.058	.004
	OR*ES	2	.776	.088	2,260	.22	.491	.058
	OR*DIR	1	.025	.002	1.035	.061	.051	.003
	RO*OR	2	14.28**	.64	31.41**	.797	1.741	.179
	ES*DIR	2	.014	.002	.438	.052	7,731**	.491
	ES*OR	4	4,329**	.52	16.29**	.803	.271	.063
	DIR*OR	2	.349	.042	6,345**	.442	1.118	.123
	OR*OR*ES	6	53,307**	.982	13.29**	.93	.682	.405
	OR*ES*DIR	3	2.39	.55	.59	.227	.069	.033
	OR*ES*DIR	6	1.34	.57	2.53	.717	.466	.318

In Table 8, when variables having an effect on type I errors of methods are examined, it is apparent that the DIF item rate has a statistically significant effect for the MH method ($F_{1,105,0.05}$ = 4.47). The main effect of the overlap rate had a significant effect on the type I error rate for MH $F_{2,105,0.05}$ = 3.14). Main effects with significant effects on the type I error of Lord's x^2 method are the DIF item rate and overlap rate ($F_{1,105,0.01}$ = 7,96) ve $F_{2,105,0.01}$ = 8,65). In addition, it can be suggested that interactions between the DIF-containing item rate and the overlap rate had a significant effect for Lord's x^2 method ($F_{2,105,0.01}$ = 4,69). In DIF obtained with MMIRT based on latent classes, factors such as reference-focal group rate, effect size and DIF item rate had a significant effect. Since one of these factors, effect size, has the highest effect value, it can be said to be more effective on type I errors $F_{2,105,0.01}$ = 101.05). Effect size and DIF item rate interaction had a significantly effect on type I errors ($F_{2,105,0.01}$ = 11.09).

When variables affecting the power of methods in Table 8 are analyzed, it can be seen that main factors such as effect size and overlap rate have a significant effect on MH method ($F_{2,105,0.01}$ = 4.41) and ($F_{2,105,0.01}$ = 41.83). When binary interactions are examined, it can be seen that interactions between reference-focal group rate and overlap rate, and also between effect size and overlap rate are statistically significant on the power of MH ($F_{2,105,0.01}$ = 14.28) and ($F_{4,105,0.05}$ = 4.33). Among triple interactions, the interaction between

Table 9

Type I Error and Power Rates According to Main Effects and Interactions Considered Significant

			Type I error			Power		
	1st condition	2nd condition	MH	Lord	MMIRT	MH	Lord	MMIRT
RO	50:50		-	-	0.168	-	-	-
	80:20		-	-	0.204	-	-	-
ES	0.5		-	-	0.35	50.16	48.5	77.65
	0.7		-	-	0.157	59.47	58.10	96.75
	1.0		-	-	0.051	64.7	63.72	100.0
DIR	20%		0.048	0.039	0.238	-	-	87.39
	40%		0.067	0.06	0.134	-	-	95.55
OR	90%		0.052	0.045	-	97.83	97.52	-
	70%		0.071	0.070	-	72.3	69.3	-
	50%		0.046	0.033	-	4.19	3.56	-
RO*OR	50:50	90%	-	-	-	100.0	100.0	-
		70%	-	-	-	87.38	85.88	-
		50%	-	-	-	5.1	4.13	-
	80:20	90%	-	-	-	95.66	95.04	-
		70%	-	-	-	57.3	52.72	-
		50%	-	-	-	3.25	3.0	-
ES*DIR	0.5	20%	-	-	0.44	-	-	68.66
		40%	-	-	0.261	-	-	86.64
	0.7	20%	-	-	0.21	-	-	93.5
		40%	-	-	0.10	-	-	100.0
	1.0	20%	-	-	0.063	-	-	100.0
		40%	-	-	0.039	-	-	100.0
ES*OR	0.5	90%	-	-	-	93.5	93.13	-
		70%	-	-	-	51.93	48.58	-
		50%	-	-	-	5.06	3.93	-
	0.7	90%	-	-	-	100.0	100.0	-
		70%	-	-	-	74.65	70.83	-
		50%	-	-	-	3.76	3.5	-
	1.0	90%	-	-	-	100.0	99.43	-
		70%	-	-	-	90.43	88.5	-
		50%	-	-	-	3.76	3.25	-
DIR*OR	20%	90%	-	0.035	-	-	98.25	-
		70%	-	0.047	-	-	96.79	-
		50%	-	0.036	-	-	76.66	-
	40%	90%	-	0.056	-	-	61.93	-
		70%	-	0.093	-	-	3.95	-
		50%	-	0.030	-	-	3.16	-

reference-focal group rate, overlap rate and effect size was statistically significant. The main effects with a significant effect on the DIF power of Lord's x^2 method are effect size and overlap rate ($F_{2,105,0.05}$ = 4.3) ve ($F_{2,105,0.01}$ = 169.6). Among binary interactions, it can be seen in Table 8 that effect size and overlap rate interaction had the most significant and highest effect ($F_{4,105,0.01}$ = 16.29). Triple interaction of reference-focal group rate, overlap range and effect size is again another variable with significant effect on power, seen via this method ($F_{6,105,0.01}$ = 13.29,). It can be interpreted that effect size had the highest main effect on the power of DIF implemented after determination of latent classes using MMIRT ($F_{2,105,0.01}$ = 40.16). Main effects and interactions found to be significant according to conditions for each method are given in Table 9.

According to Table 9, it can be said that error rates of MH and Lord's x^2 methods increase as the DIF-containing item rate increases. It can be seen that error value average is higher when the overlap rate for the MH method is 70% compared to the situation in which the overlap rate is 50%. Lowest error value averages for Lord's x^2 method can be said to emerge respectively at 50%, 90%, and 70% overlap rates. While reference-focal group rate, DIF effect size and DIF-containing item rate have a significant effect for MMIRT, the overlap rate did not create any significant differences. With an increase in effect size, the type I error rate related to MMIRT significantly decreased. This situation is also similar in condition to the DIF item rate increase. When effects of binary interactions on type I errors are analyzed, it is observed that the interaction of DIF item rate and overlap rate did not create a stable change on type I errors related to Lord's x^2 method. For MMIRT, interactions between the DIF-containing item rate and effect size were significant. When effect size and DIF-containing item rate increased in all overlap cases and group rates, it could be seen that type I error related to MMIRT decreased.

When Table 9 is analyzed, it can be seen that the power of all methods increased as the effect size increased. It can be said that, as the DIF-containing item rate increased, the only power for MMIRT increased. According to Table 9, as the overlap rate decreased, DIF performance of methods decreased. When binary interactions were examined, the power of both methods decreased when the reference-focal group rates were not equal and the overlap rate decreased. When effect size and DIF-containing item rates increased corporately, the power of DIF determined via MMIRT increased. In MH and Lord's x^2 methods, it can be seen that the power of methods increased with an increase in DIF effect size for overlap rates of 90% and 70%. This situation is slightly different at the 50% rate. With an increase in effect size, the DIF power of these methods decreased. As overlap rate decreased and number of DIF-containing items increased, the power of Lord's x^2 decreased.

Discussion

Research results have shown that, as overlap rate decreases, methods based on the manifest group become insufficient in determining DIF. When this rate falls to 50%, the power of MH and Lord's x^2 decreases. In the study, items with DIF were examined for numbers in which items were identified with DIF at A, B, and C levels within 50 repetitions. It was found that these items could not determine DIF at B and C levels when the overlap rate decreased for MH and Lord's x^2 methods. Samuelsen (2005) suggests that the MH method is able to determine DIF at the B level when the overlap rate is 70% and the effect size is 1.20, and at the C level when the overlap rate is 80% and the effect size is 1.20. Maij-de Meij et al. (2010) indicate that the difference between the error and power of Lord's x^2 method reduces in cases where the correlation between the manifest group variable and latent class variables is at .60 and higher. Finding of this study shows consistency with related studies. The overlap rate is independent of error and the power of the MMIRT method.

When the DIF-containing item rate increased, errors related to MH and Lord's x^2 increased. This finding is consistent with results found by Atalay-Kabasakal, Arsan, Gök, and Kelecioğlu (2014), Finch (2005), Stark, Chernyhenko, and Drasgow (2006), Wang and Yeh (2003). Type I error and power values related to MMIRT are affected by an increase in the number of items with DIF. De Ayala et al. (2002) have indicated in their DIF study, which was carried out on classes determined through LCA, that the power of the method increases as the DIF-containing item rate rises from 10% to 30%. Similar cases apply when a passage occurs from low effect size to higher DIF effect size level. Error toward MMIRT have substantially exceeded the acceptable error limit especially at a 0.5 effect size and 20% DIF item rate. This situation can be explained by the fact that the data does not comply with the multilevel mixture MIRT model due to the low level of DIF effect size values and the DIF-containing item rate which can reveal heterogeneity between individuals (Cho, 2007; Yüksel, 2012). When AIC values are examined, it can be concluded that the data with the 0.5 effect size and the 20% DIF rate is more suitable to the 1PL model than MMIRT model. When the DIF-containing item rate is raised up to 40%, the type I error value for DIF determined according to estimated parameters for MMIRT decreases. This situation can be explained by the fact that it is possible to reveal heterogeneous structure in data as the number of DIF-containing item increases. Cho (2007) indicates that as the number of DIF-containing items increases, RMSE and bias values related to item difficulty parameters, also AIC and BIC decrease, and the data shows a better fit to the MMIRT model. As the number of items containing DIF increases, power of MMIRT increases under all conditions. This finding also shows similarity with results obtained by De Ayala et al. (2002).

The DIF effect size was not effective on type I errors related to manifest group variable methods. As effect size increased, the power of all methods also increased.

DIF determined with estimated parameters according to MMIRT was affected by this increase most significantly. When DIF effect size increased, school level variable used in MMIRT revealed heterogeneity between individuals. Thus, data showed a better fit to the model. At a high effect size, DIF estimate through latent class variables ensured a decrease in type I errors. This finding can be said to show consistency with previous research findings (Cho, 2007; Samuelsen, 2005; Yüksel 2012).

Reference and focal group rates did not affect type I errors and power in MH and Lord's x^2 methods. According to some research findings, type I error rates obtained under equal state of reference and focal group rates tend to remain at higher levels compared with the unequal group rate (Atalay-Kabasakal et al., 2014; Erdem Keklik, 2014; Kim, 2010). There are also studies available in which type I error rates rise in the unequal group rate. Guilera, Gomez-Benito, Hidalgo, and Sanchez-Meca, (2013) indicate that the type I error rate with regard to the MH method is higher when the reference and focal group rate is less than 1 or greater than 2. The power of MMIRT is not affected by inequality of number of groups, but the type I error is lower when the number of groups is even. This finding is consistent with the finding that type I error occur more frequently with the MMIRT model in the case where groups are not equal (De Ayala et al., 2002).

An increase in the DIF-containing item rate in the unequal group rate reduced the power of manifest group methods. The finding that the power value of the MH method decreased as the rate of the DIF-containing item increase while overlap rate is 90%, shows similarities with findings by Samuelsen (2005). It can be asserted that the power value obtained for MMIRT cannot exceed 80% with the decrease in reference and focal rate, and it remains insufficient in determining DIF. On the other hand, an increase in the DIF-containing item rate increases the power of MMIRT in the unequal group rate.

Findings obtained as a result of the research show that the manifest groups cannot represent a single latent class under certain conditions. It was observed that the power of DIF determined according to the manifest group DIF dropped when the overlap rate was lower than 70%. For this reason, it may be recommended to researchers to examine the homogeneity of data first. For this purpose, data fit to a model should be assessed using information criteria (AIC and BIC). Decisions can be made based on latent classes or the manifest group in accordance with data fit to the model. When data is fit to the MMIRT model, the power of MMIRT is higher than the manifest group DIF. Therefore, when an appropriate model is used, the interpretation of items with DIF at B and C levels will be easier, and this will allow objective determination of reasons underlying item bias. MH and Lord's c^2 methods can be used in DIF when it is ensured that data consists of homogeneous groups. Since the difference between type I errors and the power of these two methods is not significant, researchers may use either of the two methods.

References

Alexeev, N., Templin, J., & Cohen, A. S. (2011). Spurious latent classes in the mixture Rasch model. *Journal of Educational Measurement*, *48*, 313–332.

Asparouhov, T., & Muthe´n, B. (2008). Multilevel mixture model. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 25–51). Greenwich, CT: Information Age Publishing, Inc.

Atalay-Kabasakal, K., Arsan, N., Gök, B., & Kelecioğlu, H. (2014). Comparing performances (Type I error and power) of IRT likelihood ratio SIBTEST and Mantel-Haenszel Methods in the determination of differential item functioning. *Educational Sciences: Theory and Practice*, *14*, 2175–2193.

Bilir, M. K. (2009). *Mixture item response theory-mimic model: Simultaneous estimation of differential item functioning for manifest groups and latent classes* (Doctoral dissertation, Florida State University, USA). Retrieved from http://diginole.lib.fsu.edu/islandora/object/fsu:182011/datastream/PDF/view

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

Cho, S.-J. (2007). *A multilevel mixture IRT model for DIF analysis* (Doctoral dissertation, University of Georgia, Athens, Greece). Retrieved from https://getd.libs.uga.edu/pdfs/cho_sun-joo_200712_phd.pdf

Cho, S.-J., & Cohen, A. S. (2007, July). *Multilevel mixture IRT model for DIF analysis*. Paper presented at the International Meeting of the Psychometric Society, The 72nd annual meeting of the Psychometric Society, Tokyo, Japan.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31–44. http://dx.doi.org/10.1111/j.1745-3992.1998.tb00619.x

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

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42*, 133–148.

Çalış, N., (2011). *Discriminant analysis based on mixture distribution models and classification* (Doctoral dissertation, Çukurova University, Adana, Turkey). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/

De Ayala, R. J., Kim, S.-H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing*, *2*, 243–276.

De Mars, C. E., & Lau, A. (2011). Differential item functioning detection with latent classes: how accurately can we detect who is responding differentially? *Educational and Psychological Measurement, 71*(4) 597–616. https://doi.org/10.1177/0013164411404221

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. http://journals.sagepub.com/doi/abs/10.1177/0146621605275728

Finch, W. H., & Finch, M. E. H. (2013). Investigation of specific learning disability and testing accommodations based differential item functioning using a multilevel multidimensional mixture item response theory model. *Educational and Psychological Measurement*, *73*(6), 973–993. https://doi.org/10.1177/0013164413494776

Frühwirth-Schnatter, S. (2006) *Finite mixture and Markov switching models.* New York, NY: Springer.

Goodman, L. (2002). Latent class analysis. In J. Hagenaars & A. McCutcheon (Eds.), *Applied latent class analysis* (pp. 3–55). New York, NY: Cambridge University Press.

Guilera, G., Gomez-Benito, J., Hidalgo, M. D., & Sanchez-Meca, J. (2013). Type I error and statistical power of the Mantel–Haenszel procedure for detecting DIF: A meta-analysis. *Psychological Methods, 18*, 553–571.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and application*. Boston, MA: Kluwer Academic Publishers Group.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. California, CA: Sage.

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.

Horst, P. (1966). *Psychological measurement and prediction*. Belmont: Wadsworth Pub. Co.

Hu, P. G., & Dorans, N. J. (1989). *The effects of deleting differentially functioning items on equating functions and reported score distributions*. Princeton, NJ: Educational Testing Service.

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 Georgia State University).

Retrieved from http://scholarworks.gsu.edu/cgi/viewcontent.cgi?article=1066&context=eps_diss

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Welsley Publishing Company.

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2010). Improvement in detection of differential item functioning using a mixture item response theory model. *Multivariate Behavioral Research, 45*, 975–999.

Mclachlan, G., & PeeL, D. (2000). *Finite mixture models*. New York, NY: John Wiley & Sons, Inc.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*(4), 297–334.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*(2), 195–215.

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

Oliveri, M. E., Ercikan, K., & Zumbo, B. (2013). Analysis of sources of latent class differential item functioning in international assessments. *International Journal of Testing*, *13,* 272–293.

Rost, J. (1997). Logistic mixture models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 449–463). New York, NY: Springer.

Samuelsen, K. M. (2005). *Examining differential item functioning from a latent class perspective* (Doctoral dissertation, University of Maryland, College Park^, USA). Retrieved from http://drum.lib.umd.edu/handle/1903/2682

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.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. London, UK: Chapman & Hall/CRC.

Stevens, P. J. (2009). *Applied Multivariate Statistics fort the Social Sciences* (5th ed.) New York, NY: Routledge Taylor and Francis Group.

Vermunt, J. K., & Magidson, J. (2005). *Technical guide for Latent GOLD 4.0: Basic and advanced*. Belmont, MA: Statistical Innovations Inc.

Yüksel, S. (2012). *Analyzing differential item functioning by mixed Rasch models which stated in scales* (Doctoral dissertation, Ankara University, Ankara, Turkey). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/