BARÚ Bartın University Journal of Faculty of Education

BUEFAD

2023, Volume 12, Issue 1, 146-166
dergipark.org.tr/buefad
DOI: 10.14686/buefad.996949

Comparison of the Methods of Examining Measurement Equivalence Under Different Conditions in Terms of Statistical Power Ratios

Research Article
Received: 17.09.2021
Revised: 15.12.2021
Accepted: 16.12.2021

***Alperen YANDI*[a]* & *Nizamettin KOÇ*[b]**

a Dr., Bolu Abant Izzet Baysal University, TÜRKİYE, https://orcid.org/0000-0002-1612-4249 * alperenyandi@gmail.com
b Prof. Dr., Maltepe University, TÜRKİYE, https://orcid.org/0000-0002-3308-7849

ABSTRACT

Validity is the most important psychometric feature that should be found in a measurement tool. Measurement equivalence is one of the evidences for the validity of measurement tools. Providing information to the researchers about the methods of identifying measurement equivalence may contribute to present a complete validity evidence. The purpose of this study is to compare the statistical power ratios of methods used for examining the measurement equivalence based on structural equation modeling and item response theory on the simulation data sets simulated by diversifying conditions of sample size, number of items, and the ratios of the items having differential item function. In accordance with this purpose, the conditions have been manipulated to include three different levels. In the analysis, multi-group confirmatory factor analysis was used which is among the methods based on the structural equation modeling, and likelihood ratio test and comparison of the item parameters methods were used which are among the methods based on item response theory. Multi-group group confirmatory factor analysis (93,50%) and likelihood ratio test (96,75%) methods have reached the highest statistical power ratio in the condition that the sample size is 1000/1000, the number of items is 40 items and the ratios of the items having differential item function is 10%. The method of comparing the item parameters (94,50%) has reached the highest statistical power ratio in the condition of sample size is 1000/1000, the number of items is 20 and the ratios of the items having differential item function are 10-20%.

Keywords: Measurement equivalence, invariance, differential item function, statistical power ratios, multi-group confirmatory factor analysis, likelihood ratio test, comparison of the item parameters method.

Ölçme Eşdeğerliğini İncelemede Kullanılan Yöntemlerin Farklı Koşullar Altında İstatistiksel Güç Oranları Açısından Karşılaştırılması

ÖZ

Geçerlilik, bir ölçme aracında bulunması gereken en önemli psikometrik özelliktir. Ölçme eşdeğerliği, ölçme araçlarının geçerliliğinin kanıtlarından biridir. Araştırmacılara ölçme eşdeğerliğini belirleme yöntemleri hakkında bilgi verilmesi, tam bir geçerlilik kanıtı sunulmasına katkı sağlayabilir. Bu çalışmanın amacı, yapısal eşitlik modellemesi ve madde tepki kuramına dayalı ölçme eşdeğerliğini incelemek için kullanılan yöntemlerin istatistiksel güç oranlarını, örneklem büyüklüğü, madde sayısı ve değişen madde fonksiyonu içeren madde oranı değişkenleri değişimlenerek oluşturulan yapay veri setleri üzerinde karşılaştırmaktır. Bu amaç doğrultusunda değişkenler üç farklı düzeyi olacak şekilde değişimlenmiştir. Analizde yapısal eşitlik modellemesine dayalı yöntemlerden çok gruplu doğrulayıcı faktör analizi, madde yanıt teorisine dayalı yöntemlerden olabilirlik oranı testi ve madde parametrelerinin karşılaştırılması yöntemleri kullanılmıştır. Örneklem büyüklüğü 1000/1000, madde sayısının 40 olduğu ve değişen madde fonksiyonuna sahip maddelerin oranları ise %10 koşulda çok gruplu grup doğrulayıcı faktör analizi (%93,50) ve olabilirlik oran testi (%96,75) yöntemleri en yüksek istatistiksel güç oranına ulaşmıştır. Madde parametrelerinin karşılaştırılması yöntemi (%94,50) örneklem büyüklüğünün 1000/1000 olması, madde sayısının 20 olması ve diferansiyel madde fonksiyonuna sahip maddelerin oranlarının 10-20% olması durumunda en yüksek istatistiksel güce ulaşmıştır.

Anahtar kelimeler: Ölçüm eşdeğerliği, istatistiksel güç oranları, çoklu grup doğrulayıcı faktör analizi, olabilirlik oranı testi, madde parametrelerinin karşılaştırılması yöntemi.

To cite this article in APA Style:
Yandı, A. & Koç, N. (2023). Comparison of the methods of examining measurement equivalence under different conditions in terms of statistical power ratios. *Bartın University Journal of Faculty of Education, 12*(1), 146-166. https://doi.org/10.14686/buefad.996949

INTRODUCTION

Identification of individual differences is one of the main objectives of the studies carried out in the field of social sciences. Individuals differ depending on the structures like environment and culture. The identification of the results obtained for individual differences is the basis of the arrangements and changes made in different fields. Environments of education, business and so on. redesigning, taking decisions at individual and group levels, and performing certain actions according to these decisions are an example of these arrangements and changes. Individual differences need to be measured accurately, therefore these arrangements and changes can be efficient. In other words, one of the most important phases in the process of the identifying the individual differences is the measurement process (Tekin, 1991). Indeed, Galton (1884) tried to identify the differences between individuals by considering several variables, and he emphasized the importance of mensuration in this process.

One of the fields in which arrangements and changes are made in line with the results of individual differences is educational sciences. Individuals' level of having psychological constructs is, also, frequently discussed in educational sciences. Since it is much more difficult to observe the psychological structures directly in this process, measurement tools are used to determine the reactions from which we can predict these psychological structures (Erkuş, 2012). Using development and adaptation studies measurement tools are introduced to the literature. The psychometric properties of the measurement tools must be adequate. Validity is the determination of the extent to which theoretical and empirical evidence is supported by interpretations based on the scores obtained in the test applications (Messick, 1989). For this reason, validity examination of the measurement tools used in researches gain more importance.

The reason for the significance of validity tests in the research process is its direct effect on the accuracy of the results. Researchers need to be cautious at some points about the validity examination. The first of these points is at which stage of the research the test of validity should be carried. Validity examination should not be carried out after the final application. It is because the researcher does not have a chance to plan on the measurement tool after the final application (Ellis & Mead, 1998). Another point that should be taken into consideration is the analysis carried out during the examination process. In most studies it is seen that certain analyses have been used in the validity examination process. When identifying validity analyses, the purpose of the study should be considered. Systematic error sources are checked during the validation examination process. The sources of systematic errors that may affect with a measurement application may differentiate. For this reason, it should be tried to predict possible sources of systematic errors according to the purpose of the research.

In a major part of the researches aimed at identifying individual differences, comparisons are made between the groups in which the individuals get involved in accordance with the variables such as gender, grade and age. In researches with this purpose, individuals in different subgroups are compared. One of the sources of systematic error that can be encountered in these researches is bias. The bias is one of the most important systematic error sources that adversely affect validity (Messick, 1995). It is expected that the individuals who are at the same level in terms of the psychological structure being measured respond at the same response level. Bias can be defined as the differentiation of the probability of individuals giving the same response (Angoff, 1993).

Types of bias that can be encountered during the measurement processes can differentiate. The concept of bias is divided into three headings as construct, method and item bias (Van de Vijver, 1998). While method and construct bias can be identified by theoretical research, item bias could be diagnosed with the help of the item response functions. One of the concepts addressed in the examination process of item bias is the measurement equivalence. Measurement equivalence can be defined as a situation in which the individuals who are at the same level in terms of latent variable but in different subgroups regarding any other variable respond to the observed variable at the same level. (Drasgow & Kanfer, 1985; Widaman & Reise, 1997).

In researches conducted in different subgroups, it is necessary to provide measurement equivalence therefore comparisons between groups can be made correctly (Reise, Widaman & Pugh, 1993; Vandenberg & Lance, 2000). However, it is seen that examination of measurement equivalence is neglected in researches in national literature on different group comparisons (Başusta & Gelbal, 2015). It can be said that the complexity of the measurement equivalence examination process, and the applications of the methods used in this process lead to this situation. For this reason, the researches about the measurement equivalence examination process are important in terms of providing information to the researchers about the use different methods by comparing applications. The researchers can be informed by clarifying the application process of measurement equivalence examination methods. This can contribute to the implementations of the measurement equivalence examination, one of the most important evidence of validity.

The methods based on the Item Response Theory (IRT) and Structural Equation Modelling (SEM) for testing measurement equivalence are priorities. In addition, latent variable models may be preferred depending on the variable properties (Güngör Culha, 2012). The researchers carry out the analyses by using the Multiple Group Confirmatory Factor Analysis (MGCFA) method based on SEM. Multi-group confirmatory factor analysis allows the analysis, based on mean-covariance structures and covariance structure, to be carried on. Some of the methods found under the IRT can be listed as Comparison of Item Parameters (CIP), Raju's Area Index, Lord's Chi-Square and Likelihood Ratio Test (LRT) (Embretson & Reise, 2000; Hambleton et al., 1991). In addition to these methods, the alligment method (Asparouhov & Muthén, 2014), which provides information about the level of measurement invariance along with the factor mean and variance between groups, and which stands out with the possibility of predicting the most appropriate models for a large number of groups, is also used in measurement equivalence examinations (Sırgancı, Uyumaz & Yandı, 2020).

Simultaneous testing of the equivalence of measurement and structural models is one of the purposes of the confirmatory factor analysis (CFA) method, one of the SEN-based methods (Asil, 2010). Equivalence of mean and covariance structures are examined in the analyses performed in the CFA process. It is suggested that covariance and mean structures are added together in models (Little, 1997; Meredith, 1993). For this reason, it is necessary to distinguish between models which deal with these two structures together and which deal only with covariance structures. The mathematical representation (Equation.1) of the covariance-based factor analytical model in multi-group confirmatory factor analysis is as follows:

$Yil = \lambda 1 \eta 1 + \lambda 2\, \eta 2 + \ldots + \lambda jm\, \eta mi + \varepsilon ji$ (Equation.1)

In this equation Yil represents the observed score of the ith person for lth item. The regression coefficients or the factor loadings of item j in the factor m, one or more factors and the random error terms having normal distribution, denote λim, ωmi, εji, respectively. In the other model, the mean values of the observed variables are also included in the model (Somer, Korkmaz, Dural & Can, 2009; Wu, Li & Zumbo, 2007). The mathematical representation (Equation.2) of the factor analytic model of this model is as follows:

$Yil = \tau j + \lambda 1(\alpha 1 + \eta 1) + \lambda 2\, (\alpha 2 + \eta 2) + \ldots + \lambda jm(\alpha m + \eta mi) + \varepsilon ji$ (Equation.2)

In this equation, unlike covariance-based multi-group confirmatory factor analysis, the factor means, denoted by αm and the intercept of item i denoted by τj are in this model. Through covariance based multiple group confirmatory factor analysis, researchers can examine equivalence of factor loadings, factor variance and covariance structures and error variances; in addition to that can examine equivalence of latent variable mean and intercepts with mean-covariance based model.

In the likelihood ratio test (Thissen, Steinberg & Gerrard, 1986), the inclusion of a differential item function (DIF) in an item is identified by comparing the likelihood ratios calculated for the two models, termed as the compact and augmented model. The parameters of item that can have DIF are released in the augmented model and calculated. In the compact model, the parameters of all the items are set to be fixed in the comparison groups. By examining the difference between the calculated 2 log likelihoods (-2 ll) for the compact and augmented modes, the item having DIF released in augmented model is interpreted. The difference between the models in the process is defined as the G2 statistic. The formula (Equation.3) used for the G2 statistic is as follows:

$G2 = -2\, (ll compact\ model - ll augmented model)$ (Equation.3)

The G2 value calculated by this formula is compared with the critical value $\chi 2$ in the degree of freedom, which is equal to the difference in the number of parameters between the two models (Thissen, Steinberg & Wainer, 1993), and examine whether the difference between the model fit statistics is significant. The fact that the difference obtained is indicative of the DIF of the released item in the augmented model.

Comparison of item parameters method which examines the parameter differences estimated for comparison groups was developed by Bock, Muraki and Pfeiffenberger (1988). In this method, analyses are performed based on the differences between the item discrimination; and item difficulty parameters were estimated for the comparison groups (Morales, Reise & Hays, 2000; Reise, Smith & Furr, 2001). Unlike other DIF identification methods, the location parameter is included in the models of this method. By using the differences between the item parameters in the analysis process, the differential item function statistic (DIFS) is calculated. The formulas used in the statistical calculation (Equation.4 and Equation.5) of differential item function are as follows (Bock et al., 1988; Morales et al., 2000; Reise et al., 2001):

Differential item function Statistic = $\hat{a}i(reference) - \hat{a}i(focal)$ (Equation.4)

Differential item function Statistic = b ̂i(reference) -b ̂i(focal) (Equation.5)

The DIFS calculated gets standardized according to the following formula (Equation.6 and Equation.7) (Muraki & Engelhard, 1989, In Thissen et al., 1993: 84).

Standardized Differential item function Statistic (SDIF)= DIFSa/(√(varâi(referreference)+varâi(focal))) (Equation.6)

Standardized Differential item function Statistic = DIFSb/(√(varb ̂i(reference)+varb ̂i(focal))) (Equation.7)

Square of the standardized differential item function statistic of each item is compared with the critical value of χ2 in one degree of freedom (Thissen et al., 1986). In this comparison, the interpretation is made as containing differential item function for the items with having SDIF is greater than the criterion χ2. If SDIF has a positive value, it is an advantage for the focus group; an advantage is given in favour of the reference group if it has a negative value.

In researches aiming to compare methods, it is important to clearly indicate in which direction the methods are compared to provide accurate information. The methods investigated in the scope of this research will be compared in terms of statistical power. The statistical power ratio in the analysis process of this research is defined as the rate of correct detection of the items containing DIF in the simulation data sets. The statistical power ratio is called "true positive" in the studies comparing measurement equivalence identification methods based on SEM and IRT and is explained in the way defined in this study (Kim & Yoon, 2011; Meade & Lautenschlager, 2004; Stark, Chernyshenko & Drasgow, 2006).

The methods compared in this study are mean-covariance based multi-group confirmatory factor analysis from SEM-based methods and comparison of item parameters method and likelihood ratio test from IRT-based methods. Providing information to researchers about methods based on different theories and making contribution to researchers about validity examination processes with the help of information provided about different methods are the reasons of the preference of these methods. In addition, multi-group confirmatory factor analysis was considered in the scope of the research because SEM-based methods were not tested on simulation data sets having different conditions.

When the literature was examined, it was determined that the statistical power and Type 1 error rates of IRT and SEM-based methods in determining DIF were discussed in different studies (Ankenmann, Witt & Dunbar,1999; Flowers, Raju & Oshima, 2002; Meade & Lautenschlager, 2004; Stark et al. ., 2006; González Roma, Hernandez & Gomez Benito, 2006; Atar & Kamata, 2011; Kim & Yoon, 2011; Kankaras, Vermunt & Moors, 2011; Elosua & Wells, 2013). The performances of the likelihood ratio test and multi-group confirmatory factor analysis methods were compared in these studies. In most of the studies, the sample sizes, and the magnitude of DIF were manipulated. It has been determined that very few studies deal with ability distributions, pattern, and DIF Type, number of items and scoring system of the measurement tool. From this point of view, the likelihood ratio test and multiple group confirmatory factor analysis methods were used in this study. In addition to these methods, comparison of the item parameters method, which is easier to apply, is also used The conditions of ratio of item having DIF, which was less discussed in the studies, was included in the study as a simulation condition. In addition, the condition of the number of items is another condition that changed in the study. Considering the real application conditions, it is possible to encounter the simulation of the number of items in the measurement tools that can be used. It is thought that the inclusion of these two conditions in the study will provide new information to the researchers. In addition, presenting results regarding comparison of the item parameters method, which is a highly useful method, will enable researchers to obtain information about this method and to obtain evidence regarding the performance of this method under different conditions.

In summary, when the literature was evaluated, it was seen that different conditions were manipulated and different methods were tried in the process of determining measurement equivalence. In addition, it has been determined that the findings in the studies and the conditions examined vary. In this study, different conditions were manipulated based on the literature and CFA and IRT based methods were compared. The general purpose of this research is to compare measurement equivalence identification methods that can be used for multiple scored data under different conditions. For this purpose, evidence has been found that the most appropriate measurement equivalence identification method for different conditions that may be encountered in real applications. Results obtained for identification of measurement equivalence by perform MGCFA, CIP and LRT on multiple scored simulation data sets simulated with manipulating conditions of sample size, number of items, and ratios of items having DIF were compared. For this purpose, the following questions were answered in the survey:

1. How do the statistical power ratios in the measurement equivalence analyses with

a. Multi group confirmatory factor analysis,

b. Likelihood ratio test,

c. Comparison of item parameters

on the simulation data sets with the conditions of different sample sizes, different number of items, and different ratios of items having DIF differ?

2. What kind of differentiation arises when the statistical power ratios of the three methods used are compared?

METHOD

Research Design

In this study, it is aimed to compare the statistical power ratios of measurement equivalence testing methods on simulation data sets simulated in different conditions and to provide the researchers with information about measurement equivalence process. The research is basic research in that it offers contribution to the theoretical studies. Karasar (2005) stated that the main purpose of the basic researches is to "bring up to date to existing information".

Generating The Data Sets

In this section, the characteristics of the simulation data sets used in the research are explained. The simulation data sets were simulated by manipulating the conditions identified within the scope of the research. Some conditions are fixed for subgroups. 100 replications were made for all data sets. Table 1 shows the conditions manipulated and fixed.

Table 1. Conditions used in the simulation data sets generation process.

Manipulated conditions	Fixed conditons
Sample size	Number of response category
Number of items	Item Response Theory Model
Ratios of item having DIF	DIF Type
	Magnitude of DIF
	Ability distribution of individuals
	Number of groups (2 subgroup)

Manipulated conditions

The manipulated conditions for datasets, such as sample size, number of items, and ratios of item having DIF are explained in this section. Each condition is split into three sub-categories. Therefore, in the data sets, 27 (3x3x3) conditions were consisted of.

Sample Size

In studies conducted using measurement equivalence methods, it has been indicated that sample sizes are in the range of 400-2000 (Atalay, Gök, Kelecioğlu & Arsan, 2012; Bolt, 2002; Holmes Finch & French, 2007; Huang, Church & Katigbak, 1997; Kazelskis, Thames & Reeves, 2004; Kim & Cohen, 1998; Korkmaz, 2005; Nachtigall, Kroehne, Funke & Steyer, 2003; Narayanan & Swaminathan, 1994; Reise, Wideman & Pugh, 1993; Somer et al., 2009; Somer, 2004). Considering these results in the literature and actual application conditions, in this study the sample size variable manipulated as 500 (250/250), 1000 (500/500) and 2000 (1000/1000).

Number of items

The number of items is discussed as another variable influencing the methods used in the measurement equivalence identification researches. In studies conducted using actual or simulation data sets that can be accessed in the literature, it is seen that the item numbers change between 6 and 180 items (Clauser & Mazor, 1998; Dodeen, 2004; Kazelskis et al., 2004; Korkmaz, 2005; Atalay et al., 2012). In the studies, it is observed that the number of items variables differ in a wide range. It has been regarded as a necessity to manipulate this variable in this study due to the effect of the number of items variable on the statistical power of the methods to identify measurement equivalence and the use of measurement tools consisting of items at different numbers in actual applications. In this study, simulation data sets of 20, 40 and 60 items were produced in order to correspond with the conditions that may be encountered in actual applications.

Ratios of items having DIF

The items having DIF affect the validity of the test negatively, because DIF is a source of bias (Camilli & Shepard, 1994). In addition, the items having DIF negatively affect the power of the methods used in the DIF analysis (Fidalgo, Mellenbergh & Muniz, 2000; Wang & Yeh, 2003; Holmes Finch & French, 2007; Atalay et al., 2012). When the studies are examined, it is seen that the ratios of the items having DIF are in the range of 0% and 40% in simulation data sets studies (Rogers & Swaminathan, 1993; Narayanan & Swaminathan, 1994; Fidalgo et al., 2000; Holmes Finch & French, 2007). In line with this information, for the simulation data sets simulated in this study, ratios of the items having DIF have been manipulated at ratios of 10%, 20% and 30%. According to these proportions, each data set contains different number of items having DIF belong to number of total items. The items having DIF are in the first part of the data sets. For 10%, 20% and 30% ratios, first 2, first 4 and first 6 items in data sets consisting of 20 items; first 4, first 8, and first 12 items in the data sets consisting of 40 items; first 6, first 12 and first 18 items in the data sets consisting of 60 items have DIF respectively.

The aim of this study is to examine the statistical power of the DIF analysis method based different theoretical models. The definition of the statistical power in this study is the rate of correct detection of the items containing DIF in the simulation data sets. Because of this definition, the no DIF items were not included in the data sets. No DIF items can be included in the studies which aimed to examine type I error. A Type I error is the incorrect rejection of a true null hypothesis. If the DIF detection methods describe DIF on the no DIF items, type I error can be examined.

Variables of fixed

The statements about the variables that are fixed in data sets appear in this section. The fixed variables are number of response category, Item Response Theory Model , the DIF Type, magnitude of DIF and ability distribution of individuals.

Number of response category

In this study, the data sets have the items including five-responding category It has been decided that at least four of the responding category numbers are taken in simulation data sets (Dodeen, 2004; Garrett, 2009; Bilican Demir, 2014). Also, the scales having the five-response category were commonly applied in the researches. Therefore, five-response category is selected for the adaptation of the real application

Item Response Theory Model

Samejima's (1969) Graded Response Model was used to simulate response patterns in simulation data sets. This model is preferred because it has mathematical models for ordinal and multiple scores, therefore it is suitable for measuring the characteristics that can be measured with multiple scoring such as performance and attitude (Samejima, 1997).

DIF Type

In a Graded Response Model, items may differ in terms of threshold and discriminant parameters in the subgroups. In this case, a uniform and non-uniform differential item function can occur for the items. In this study, cases were used items having uniform DIF. Therefore analyses are performed only on the cases where the items differ only in terms of threshold parameters.

Magnitude of DIF

In this study, the change between the parameters of 0.49 logit units identified by the Educational Testing Service for the moderate change level was taken as the quantity of the DIF.

Ability distribution of individuals

Roussos and Stout (1996) and Tian (1999) emphasized that the difference between 0.50 and 1.00 standard deviations between groups' ability distribution averages is close to real applications, based on real test applications and expert opinion. In this study, the ability distributions of the subgroups were limited to the ability distributions (N (0, 1)) with a mean of 0.00 and a unit normal distribution characteristic of 1.00 standard deviations.

Analysis of Data Sets

The analysis was carried out on the simulation data sets simulated different conditions by using multi-group confirmatory factor analysis method based on SEM, likelihood ratio method and comparison of item parameters methods based on IRT.

In this study, measurement equivalence analysis was carried out for the MGCFA method by following the stepwise test method suggested by Meredith (1993) Suggested steps for testing the measurement equivalence are:

- Configural invariance: The basic step in testing the measurement equivalence. At this step, it is examined whether the measurement tools applied to different groups have the same factor structure in each group.
- Metric invariance: In the step of metric invariance, the factorial loadings (λ) of the items in the measurement tool are tested for equivalence in the groups being measured (Vandenberg & Lance, 2000).
- Scalar invariance: In the step of scalar invariance, the regression intercepts are fixed for different groups in addition to the factor structure in the step of configural invariance and the factor loadings in the step of metric invariance (Brown, 2015; Hong, Malik & Lee., 2003).
- Strict invariance: After the steps of configural, metric and scalar invariance are confirmed, the error terms (ε) related to the items are fixed in final step. It is checked whether the observed variables have the same amount of error terms in the groups participating in the measurement application in the step of strict invariance.

In MGCFA performed by using LISREL 8.72 program. In the analysis process, maximum likelihood estimation was used because all data sets were produced correspondent with normal distribution. And in all steps, covariance matrixes were used in the analysis due to same reason (Jöreskog and Sörbom, 1999).

MULTILOG 7.03 (Thissen, 1991) program was used in the analysis of likelihood ratio test. The data sets simulated for each condition was adapted to the analysis and the program as a format. For each data sets in folders, the compact model syntax file is written first. Then, the augment model syntax files are prepared for each item. Results were obtained by running the syntax files for the data sets in each folder. The -2 loglikelihood values were deduced from each analysis obtained. Then the differences between these values for each data set is calculated by using the excel program. Whether or not the item has a differential item function has been identified by comparing these differences with critical χ2 value for degrees of freedom specified.

In the analysis process, comparison of the item parameters method was carried out by using the PARSCALE 4.1 program (Muraki & Bock, 1991). The data sets simulated for each condition was adapted to the analysis and the program as a format as such in LRT analysis. One syntax file has been created for each of the data sets. The outputs of the analysis are organized. And then tables are created to indicate the significance of the difference of each item parameter were taken. By using the Excel program, the items in which the difference of parameters are significant are identified in these tables and decided for the items whether having DIF or not.

Research Ethics

The research was carried out on simulative data; it does not contain any application made on individuals in any way.

FINDINGS

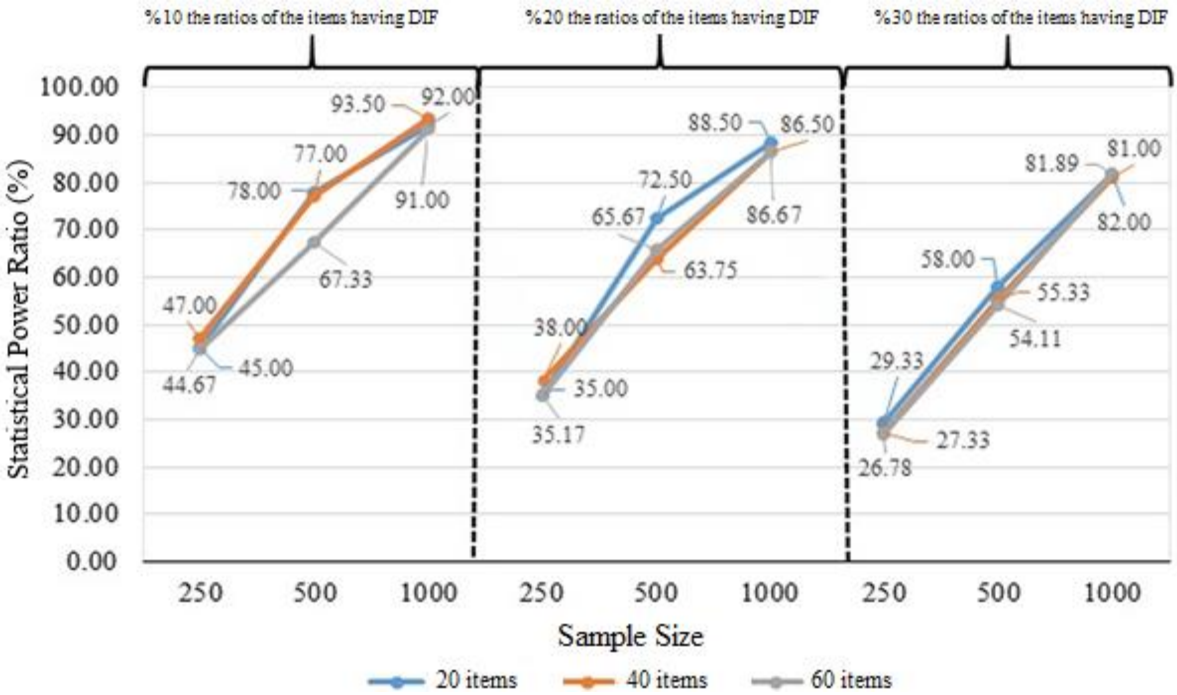
In this section, the results of the analysis are presented with graphics. Tables related to the analysis results are in the appendices. Graphics are preferred in the text therefore the results can be explained simply and clearly.

Findings Related To Statistical Power Ratios of MGCFA

The findings and interpretations of the statistical power ratios obtained from the analysis carried out by MGCFA are included below this title. The statistical power ratios of the MGCFA are given in Appendix 1.

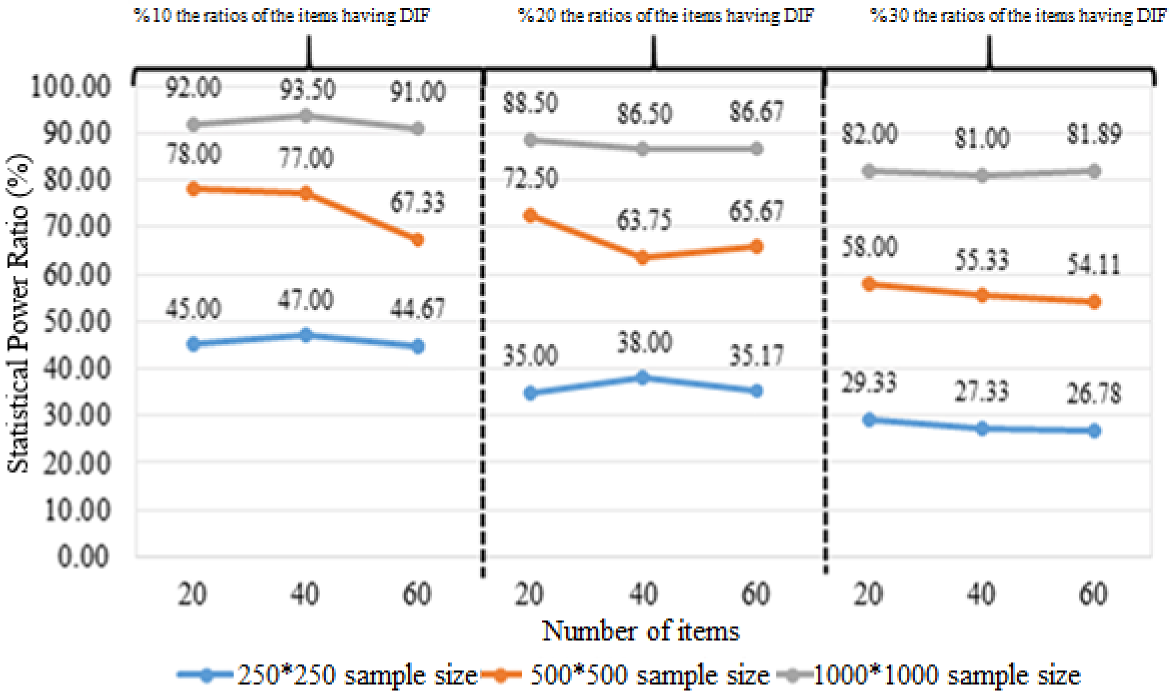
According to Appendix 1, statistical power ratios of MGCFA are in the range of 26.78% and 93.50%. The lowest power ratio for this method was calculated when the focus and reference group sample size was 250, the number of items was 60, and the ratios of the items having DIF was 30%. The highest statistical power ratio was calculated for the condition that the focus and reference group sample size was 1000, the number of items was 40, and the ratios of the items having DIF was 10%.

The results of the change in the statistical power ratios according to the sample size variable are presented in Graph 1.



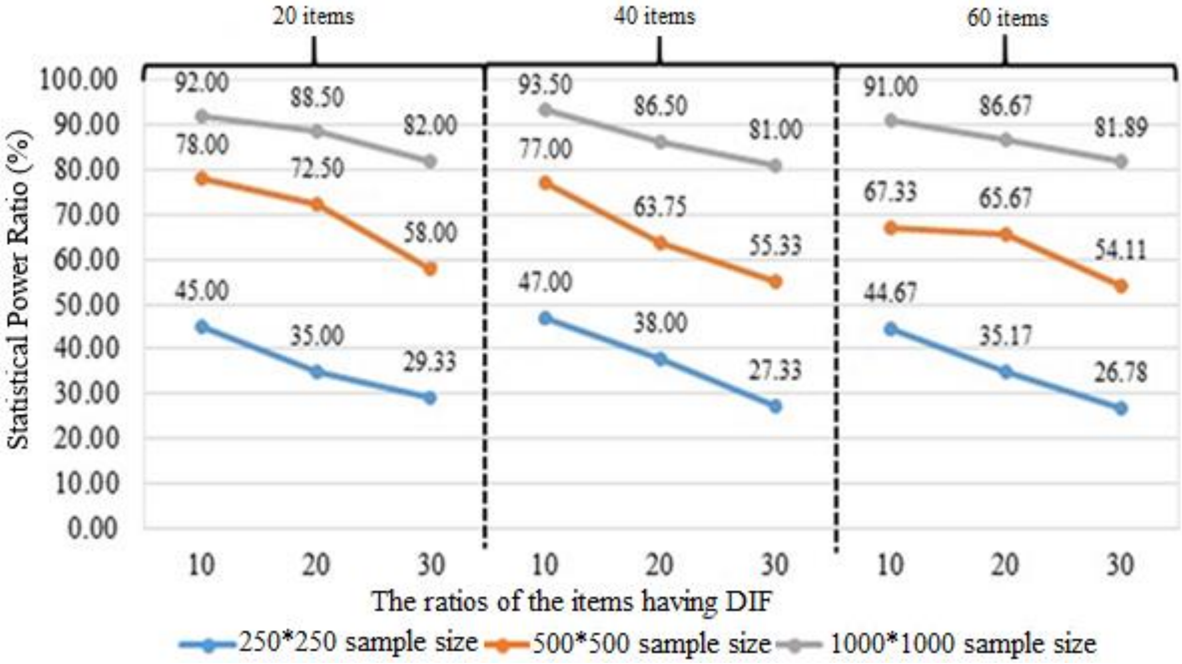
Graph 1. *Statistical power ratios for variable of sample size (MGCFA).*

According to the findings presented in Graph 1, the statistical power ratios obtained for the MGCFA increase belong to increment of the sample size at all levels of the variable number of items and ratios of the items having DIF. The results obtained for viewing the change in the statistical power ratios according to the item number variable are presented in Graph 2.



Graph 2. *Statistical power ratios for variable of number of items (MGCFA).*

According to the findings presented in Graph 2, the statistical power ratios of MGCFA decrease in low amounts depending on the number of items. The results presented in Graph 3 are used to illustrate the statistical power ratios given in Appendix 1 for variable of the ratios of the items having DIF.



Graph 3. *Statistical power ratios for variable of the ratios of the items having DIF (MGCFA).*

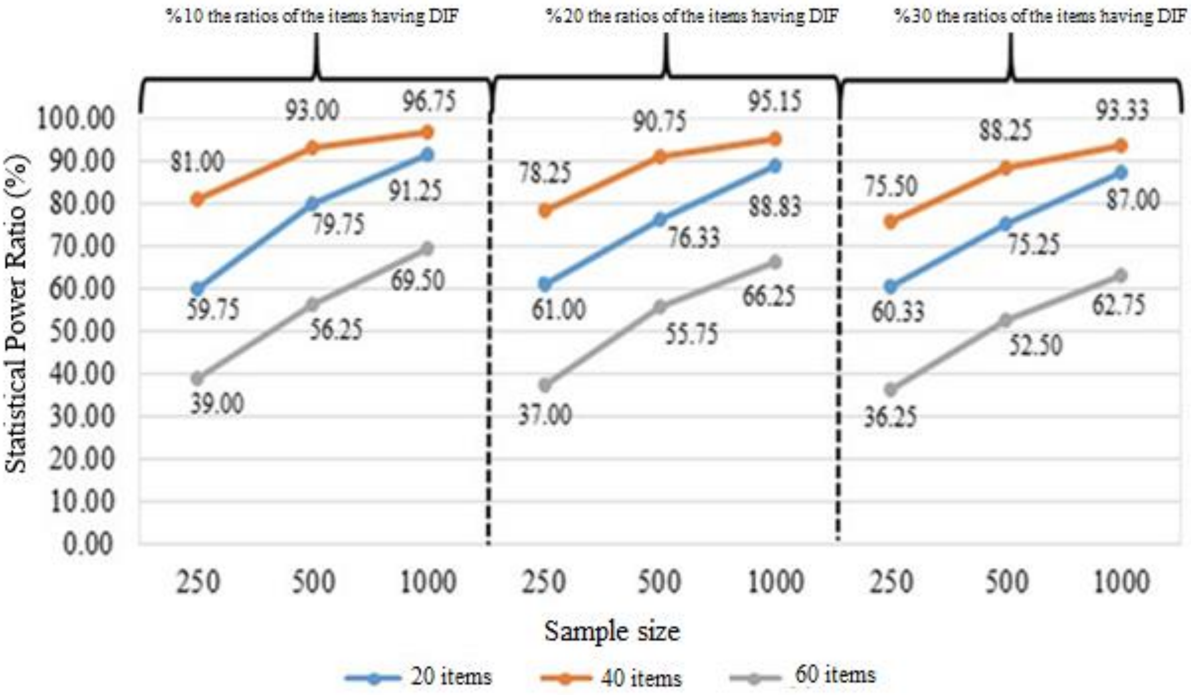
The statistical power ratios of the MGCFA are reduced by increasing the variable of ratios of the items having DIF according to the values presented in Graph 3.

Findings Related to Statistical Power Ratios Of LRT

Findings and interpretations on the statistical power ratios obtained by the analysis of the likelihood ratio test are included below this title. The statistical power ratios of the multiple LRT are given in Appendix 2.

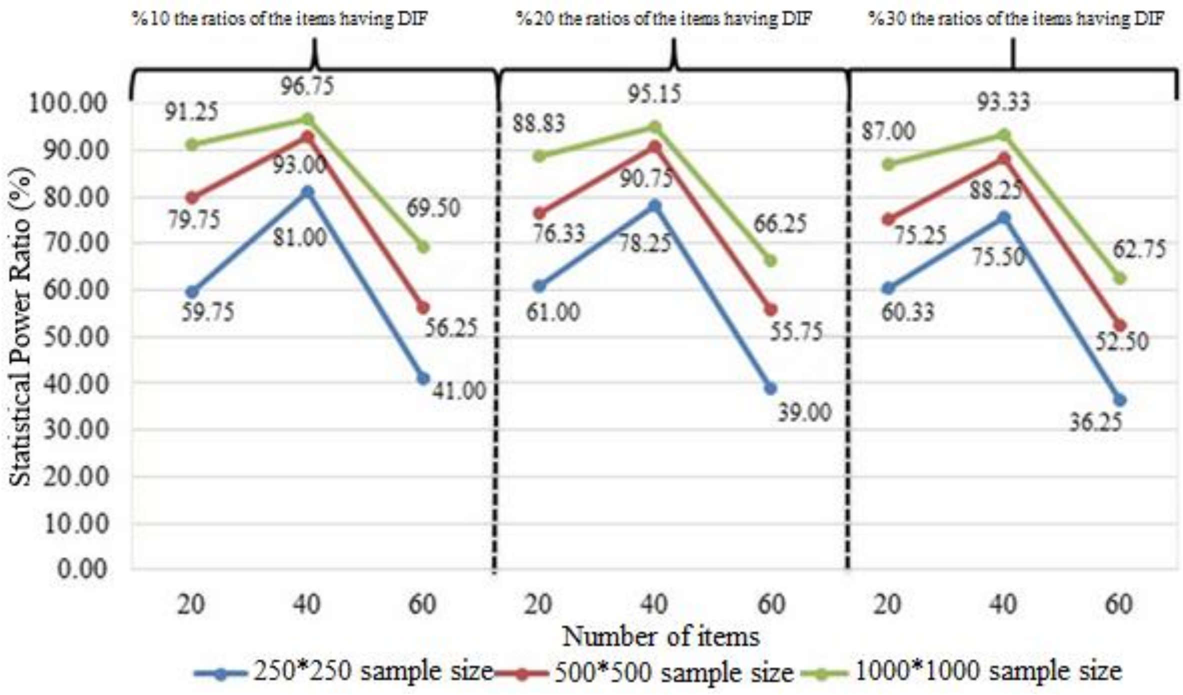
The results presented in Appendix 2 indicate that the likelihood ratio test statistical power ratios are in the range of 36.25% and 96.75%. The lowest power ratio for this method was obtained when the focus and reference group sample size were 250, the number of items was 60, and the ratios of the items having DIF was 30%. The highest statistical power ratio is calculated for the condition that the focus and reference group sample size was 1000, the number of items was 40, and the ratios of the items having DIF was 10% as is the case with MGCFA.

The results are presented in Graph 4 therefore the change in the statistical power ratios can be seen according to the sample size variable clearly.Findings section can include the description of your statistics or findings.



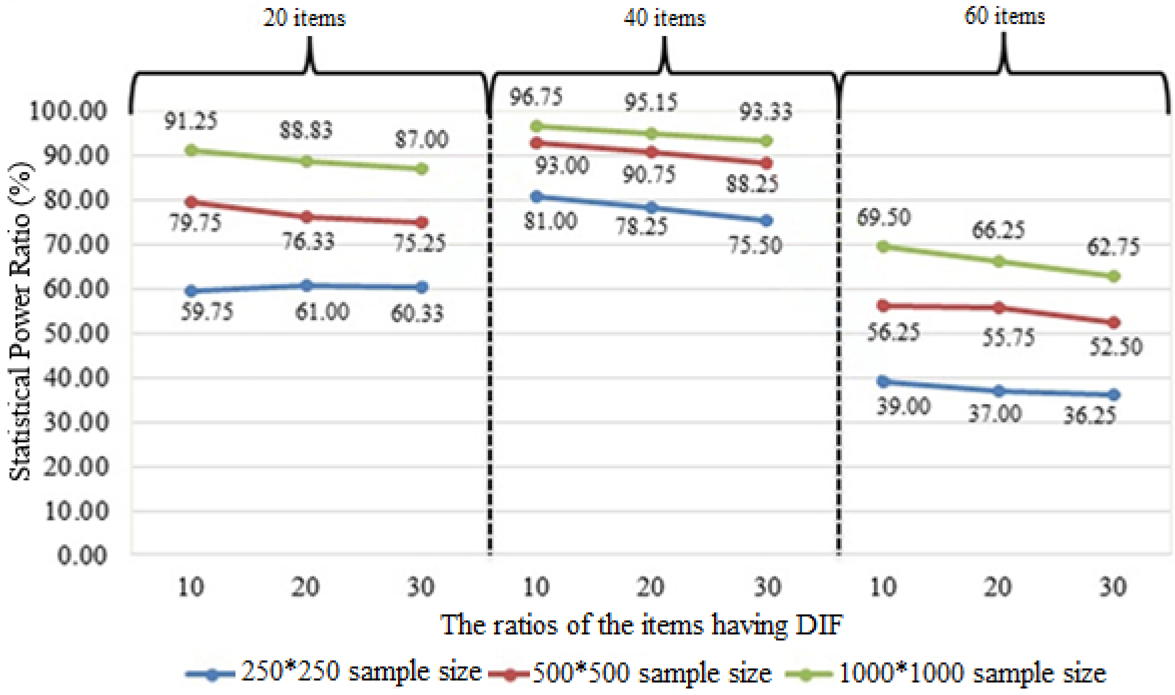
Graph 4. *Statistical power ratios for variable of sample size (LRT).*

When the curve in Graph 4 is examined, it is seen that the statistical power ratios obtained for the LRT increase due to increment of the sample size. The statistical power ratios obtained for the variable of number of items in the analyses by using the likelihood ratio test are given in Graph 5.



Graph 5. *Statistical power ratios for variable of number of items (LRT).*

The findings presented in Figure 5 indicate that high statistical power ratios for 40-item data sets are available for LRT. In other words, it reveals that the number of the ideal number of items for LRT is 40 for the data sets simulated in this study. The statistical power ratios for the variable rate of the items having DIF are given in Graph 6.



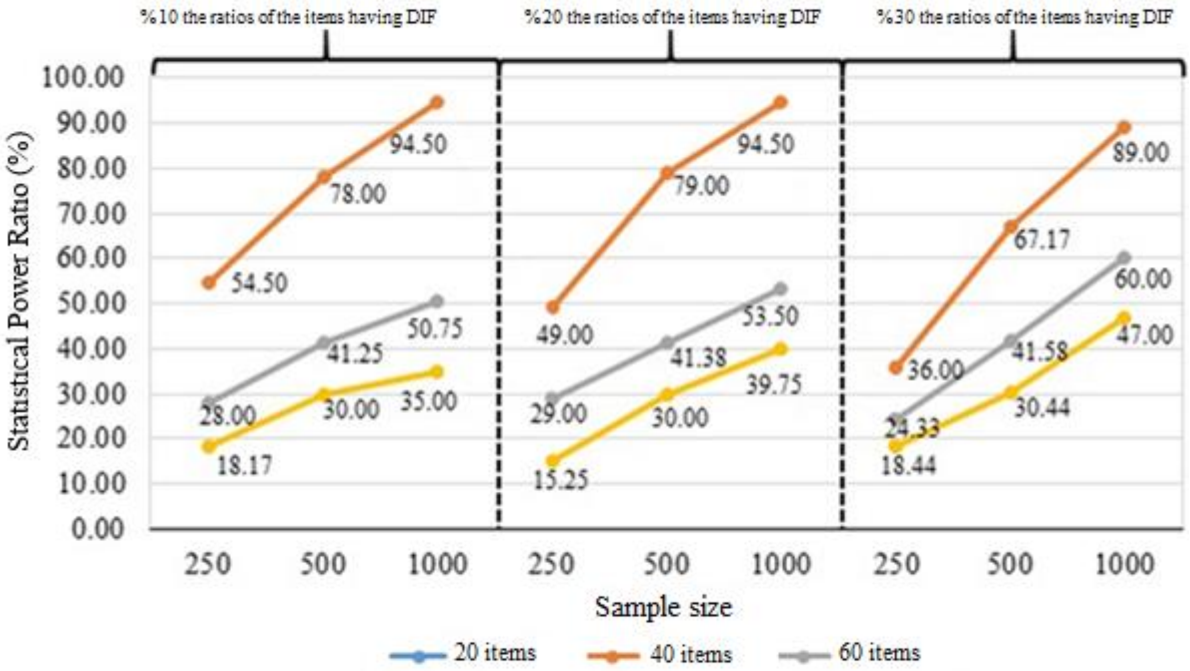
Graph 6. *Statistical power ratios for variable of the ratios of the items having DIF (LRT).*

When Graph 6 is examined, increment of the ratios of the items having DIF causes to a slight decrease on the statistical power of LRT. However, this is not valid for every condition presented in Graph 6 where it could be seen in the data sets having 250*250 sample size and consisting of 20 items.

3.3. Findings related to statistical power ratios of CIP

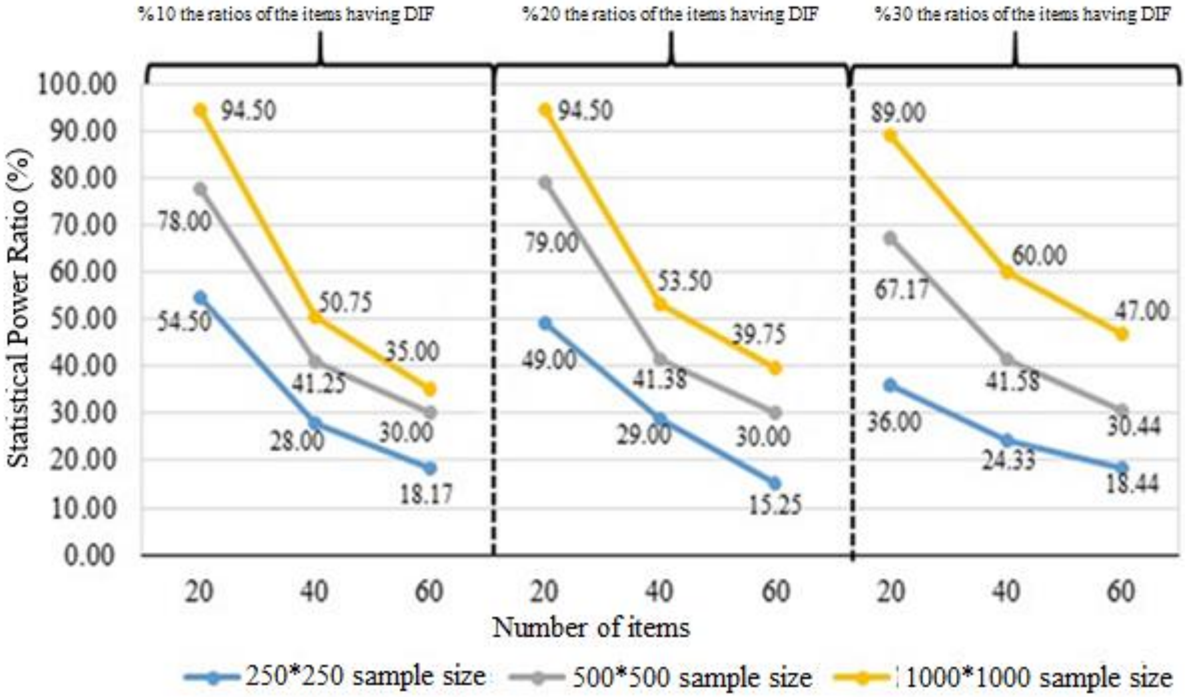
In this section, findings and interpretations related to the statistical power ratios obtained in the analyses made by CIP are given. The statistical power ratios of the method of item interpretation parameters are given in Appendix 3.

The results presented in Appendix 3 show that the likelihood ratio test's statistical power ratios are in the range of 15.25% and 94.50%. The lowest power ratio for this method was obtained when the focus and reference group sample size was 250, the number of items was 60, and the ratios of the items having DIF was 20%. The condition where the highest statistical power ratio is calculated is that the focus and reference group sample size is 1000, the number of items is 20 and the ratios of the items having DIF are 10 and 20%. The statistical power ratios for the sample size variable are given in Graph 7.



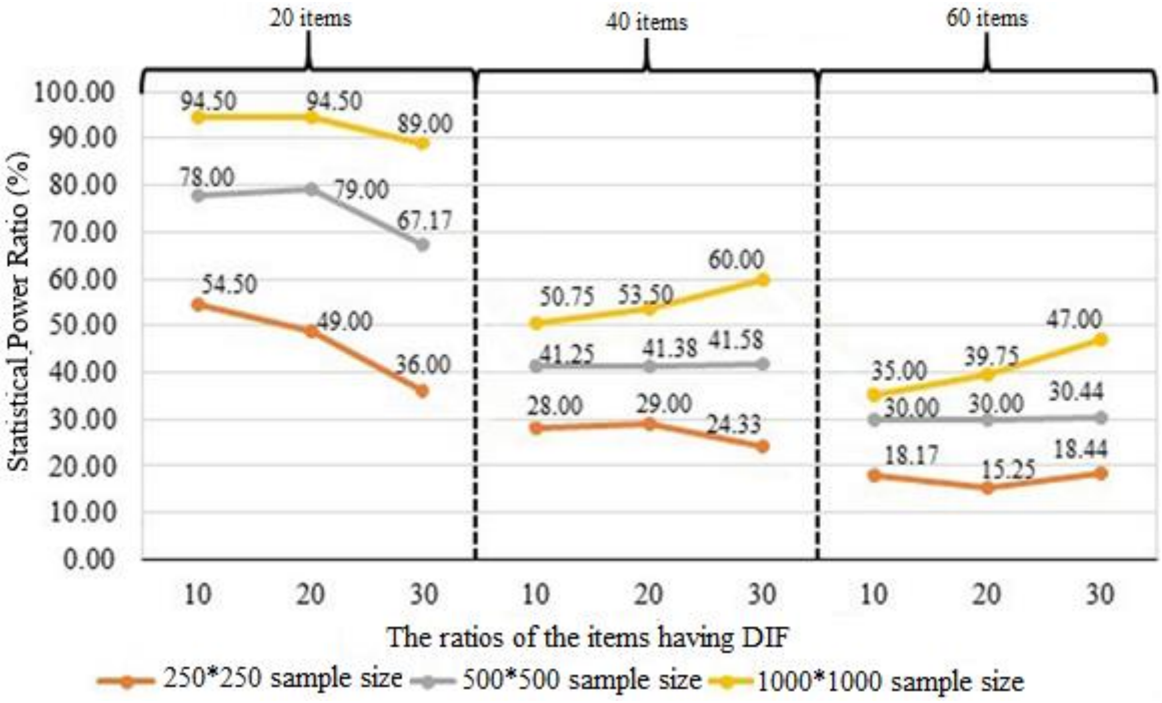
Graph 7. *Statistical power ratios for variable of sample size (CIP).*

Graph 7 shows that when the sample size increases, the statistical power ratios obtained by using CIP increase. The statistical power ratios for the variable number of items are given in Graph 8.



Graph 8. *Statistical power ratios for the variable number of items (CIP).*

When Graph 8 is examined, it is seen that the statistical power ratios obtained by using CIP decreases when the number of items increases. The statistical power ratios are presented in Graph 9 for the variable of the ratios of the items having DIF.



Graph 9. *Statistical power ratios for variable of the ratios of the items having DIF (CIP).*

When Graph 9 is examined, it is understood that the variable of the ratios of the items having DIF does not have a clear effect on the statistical power ratios of CIP. When the statistical power ratios are examined, the decrease in the data sets consisting of 20 items; for the data sets consisting of 40 and 60 items, both increase and decrease cases are seen due to the increase in variable of the ratios of the items having DIF.

Findings Related to Common Effect

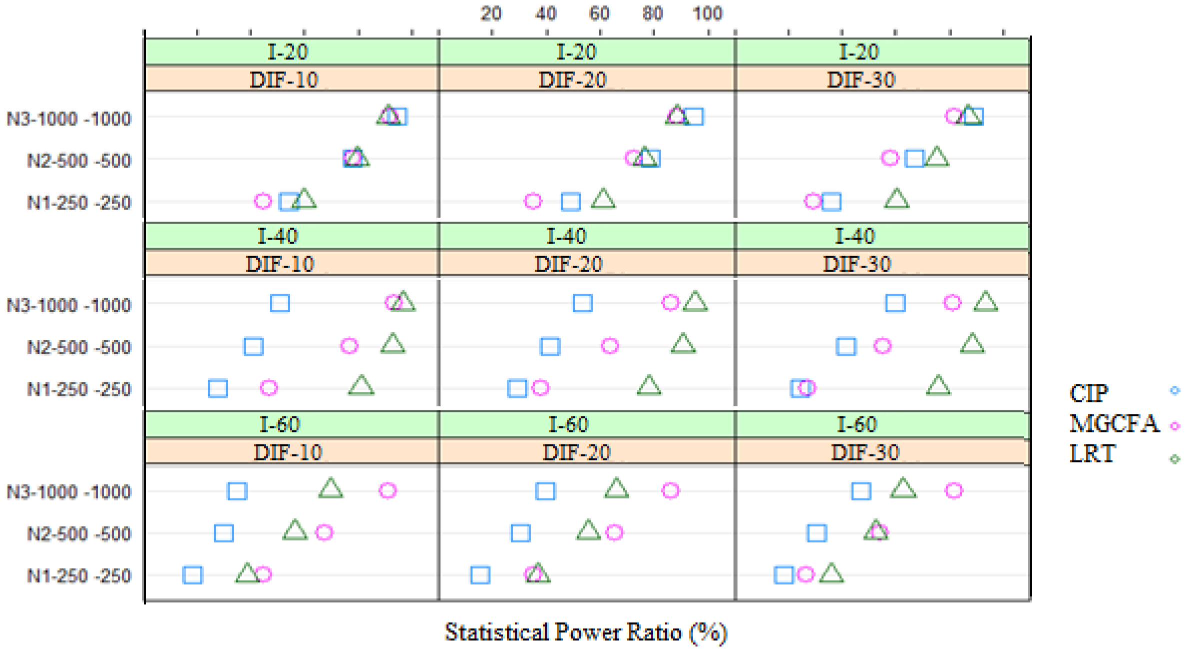
The statistical power ratios of the methods calculated under the manipulated conditions in this study are presented together in Appendix 4. The results in Appendix 4 were reviewed and interpretations were made about conditions in which the statistical power ratios of the methods are high.

When Appendix 4 is examined, it can be seen that LRT gives higher results with the exception of two conditions in which the focus and reference group sample sizes are 250, 60 items –10% the ratios of the items having DIF and the focus and reference group sample sizes are 500. It was identified that the statistical power ratios of the methods differ according to the item number variable in the condition that the sample size condition is 1000/1000. It provides clues about using CIP for data sets consisting of 20-items, LRT for data sets consisting of 40-items, and MGCFA method for data sets consisting of 60-items in the 1000/1000 sample size.

The conditions in which the level of the statistical power ratios of the methods reached the highest level for simulation data sets simulated in the study are arranged as:

- 1000/1000 sample size, consisting of 40 items, %10 the ratios of the items having DIF for MGCFA (%93.50),
- 1000/1000 sample size, consisting of 40 items, %10 the ratios of the items having DIF for LRT (%96.75),
- 1000/1000 sample size, consisting of 20 items, %10 and %20 the ratios of the items having DIF for CIP (%94.50).

Graph 10 represent the common effects of conditions manipulated on the statistical power ratios of the methods.



Graph 10. *Common effects of conditions manipulated on methods*.

When Graph 10 is examined, it is seen that LRT gives higher statistical power ratio than the other two methods for the 16 of the 27 different conditions The MGCFA and the CIP gave higher results in 7 and 4 conditions than the other two methods respectively.

DISCUSSION & CONCLUSION

The findings of the sample size variable obtained from the MGCFA in this study correspond with the findings of Stark et al. (2006), González Roma et al. (2010), Kim and Yoon (2011), Kankaras et al. (2011) and Wells (2013). In this study, like other studies, the increase in sample size was positively reflected in the statistical power ratios of MGCFA. Meade and Lautenschlager (2004) reported that, unlike the findings obtained, the increment of sample size caused a decrease in the statistical power ratios for the MGCFA. It can be argued that this difference arises from the differences between the conditions addressed in the studies. In the study conducted by Meade and Lautenschlager (2004), it is stated that the maximum statistical power ratio for the MGCFA is obtained when sample size is 150, 500 - 1000 for each group is 150. The statistical power ratios for the 20-40 and 60-item data sets are in the range of 29.33 - 92.00; 27.33 - 93.50 and 26.78 - 91.00 respectively. When the other variables are fixed, there are nine different cases for 20, 40, 60 item data sets. The highest statistical power ratios were obtained in six of these nine cases, having 20 items and three of these nine cases, having 40 items. In these findings, in the data sets simulated within the scope of the study, it is revealed that the ideal statistical power ratios were calculated for MGCFA as a result of the analyses, carried out in the 20-item data sets. Besides statistical power of MGCFA decrease belong to the increment of the ratios of the items having DIF. These findings correspond with the study conducted by Kankaras et al. (2011), but not with the Meade and Lautenschlanger (2004). In the study conducted by Kankaras et al. (2011), one and three items in the five-item were modified to affect the equivalence adversely in the simulation data sets. When the number of items having DIF was 1, the statistical power ratio was 56.50%, but when it was 1, it was identified to be 55.90% in the analyses made by using MGCFA. In the study by Meade and Lautenschlager (2004), two or four items from the six items constituting the data sets were simulated to include DIF. It has been reported that, in the analyses performed on the data sets by the MGCFA, the number of items having DIF did not cause a significant change in the statistical power ratios. Since the number of items in the data sets is limited to six in this study, it is possible to say that this difference between this study and Meade and Lautenschlager (2004) arises because of simple structure of the model used in analysis.

In this study, the findings of variable of the sample size for the LRT correspond with the findings of Ankenmann et al. (1999), Meade and Lautenschlager (2004), Stark et al. (2006), Atar and Kamata (2011), Kankaras et al. (2011), Kim and Yoon (2011) and Elosua and Wells (2013). In these studies, values between 100 and 2000 were chosen for the sample size. In all of these studies it has been found that the statistical power ratios of the LRT also change in parallel with the change in sample size, as identified in this study. Another conclusion regarding LRT is related to variable of the number of items. It was identified that the ideal number of items that can be used for LRT in the simulation data sets simulated in this study. In other words, the statistical power ratios of LRT reached the highest level in data sets consisting of 40 items. The results for the variable of the ratios of the items having DIF differ from the results obtained by Meade and Lautenschlager (2004) while correspond with the results obtained by Kankaras et al. (2011). In this study, it was identified that the increment of the variable of the ratios of the item having DIF resulted in a decrease in the statistical power ratios of LRT. However, Meade and Lautenschlager (2004) reported that this variable did not have a significant effect on the statistical power ratios of the LRT. It can be argued that this difference of the results in the studies originated from the differences in the manipulated conditions and the models used in the analyses.

The statistical power ratios of the CIP reached the highest level when the sample sizes of the reference and focal group were 1000/1000 in all levels of the variable of the number of items and the variable of the ratios of the items having DIF. In addition, the results obtained for this method indicate that when the ratios of the items having DIF is kept constant, the increment of the number of items affects the statistical power ratio negatively. It was found that the highest statistical power ratios were obtained for the 20 item data sets when the variables of the sample size and the ratios of the items having DIF were kept constant.

Recommendations:

In this section, recommendations for the application of the methods, whose statistical power ratios were examined are presented.

- It may be advisable to use the likelihood ratio test for the measurement equivalence analysis in the studies to be performed if the sample size of the comparison groups is between 250 and 500
- When a measurement tool consisting of 40 items with one dimension is used, it may be advisable for researchers to prefer the likelihood ratio test for measurement equivalence analysis.
- When a measurement tool consisting of 60 items under one dimension is used, it may be advisable for researchers to prefer the multi group confirmatory factor analysis for measurement equivalence analysis.
- In cases where the sample sizes of the focus and reference groups are approximately 1000 and the data sets are consisting of 20 items, comparison of the item parameters method can be used to examine measurement equivalence.
- It may be advisable for the researchers to plan each sample size of the focus and reference group to be at least 500 therefore the likelihood ratio test can reach the sufficient level for the statistical power ratio (> 70%) in cases where the data sets are consisting of 20 and 40 items.
- In order that the multi-group confirmatory factor analysis method can reach the statistical power ratios at a sufficient level regardless of the number of items, it can be suggested that the sample size of the comparison group should be approximately 1000.
- It may be advisable for the researchers to plan each sample size of the comparison groups to be at least 1000, therefore the comparison item parameters method can reach sufficient level for the statistical power ratio in cases, where the data sets are consisting of 20.

Statements of Publication Ethics

The research was carried out on simulative data; it does not contain any application made on individuals in any way.

Researchers' Contribution Rate

The article was produced from the first author's doctoral thesis. The study process was carried out by the first author, under the supervision of Prof. Dr. Nizamettin KOÇ.

Conflict of Interest

There is no conflict of interest in the study.

REFERENCES

Angoff, W.H. (1993). Perspectives On Differential Item Functioning Methodology. Holland and Wainer (Ed.), *Differential Item Functioning* içinde (s. 3-23). Lawrence Erlbaum Associates, Publishers, New Jersey.

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An Investigation of the Power of the Likelihood Ratio Goodness-of-Fit Statistic in Detecting Differential Item Functioning. *Journal of Educational Measurement*, *36*(4), 277-300. Retrieved from https://doi.org/10.1111/j.1745-3984.1999.tb00558.x

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

Atar, B., & Kamata, A. (2011). Comparison of IRT likelihood ratio test and logistic regression DIF detection procedures. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *41*(41).

Bock, R. D., Murakl, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, *25*(4), 275-285.

Bolt, D. M. (2002). A Monte Carlo Comparison of Parametric and Nonparametric Polytomous DIF detection methods. *Applied Measurement in Education*, *15*,113-141. Retrieved from https://doi.org/10.1207/S15324818AME1502_01

Camilli, G., Shepard, L. A., & Shepard, L. (1994). *Methods for identifying biased test items (Vol. 4).* CA: Sage.

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. Retrieved from https://doi.org/10.1111/j.1745-3992.1998.tb00619.x

Dodeen, H. (2004). Stability of differential item functioning over a single population in survey data. *The Journal of experimental education*, *72*(3), 181-193. Retrieved from https://doi.org/10.3200/JEXE.72.3.181-193

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology*, *70*(4), 662. Retrieved from http://dx.doi.org/10.1037/0021-9010.70.4.662

Ellis, B., & Mead, A. (1998). Measurement equivalence of a 16PF Spanish translation: An IRT differential item and test functioning analysis. *In 24th meeting of the International Association of Applied Psychology*, San Francisco.

Elosua, P., & Wells, C. S. (2013). Detecting DIF in polytomous items using MACS, IRT and ordinal logistic regression. *Psicológica*, *34*(2)

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Erkuş, A. (2012). *Psikolojide ölçme ve ölçek geliştirme*. Ankara: Pegem Akademi Yayınları.

Fidalgo, A. M., Mellenbergh G. J., & Muniz, J. (2000). Effects of Amount of DIF, Test Length, and Purification Type on Robustness and Power of Mantel-Haenszel Procedures. *Methods of Psychological Research Online*. *5*(3), 43-53.

Flowers, C.P., Raju, N.S., & Oshima, T. C. (2002). A comparison of measurement equivalence methods based on confirmatory factor analysis and item response theory. *Paper presented at the Annual Meeting of the National Council on Measurement in Education*, New Orleans.

Galton, F. (1884). Measurement of character. *Fortnightly Review*, *36*, 179-185.

González Roma, V., Hernandez, A., & Gomez-Benito, J. (2006). Power and Type I error of the mean and covariance structure analysis model for detecting differential item functioning in graded response items. *Multivariate Behavioral Research*, *41*(1), 29-53. Retrieved from https://www.tandfonline.com/doi/abs/10.1207/s15327906mbr4101_3

Güngör Culha, D. (2012). *Örtük sınıf analizlerinde ölçme eşdeğerliğinin incelenmesi*. [*Investigating measurement equivalence with latent class analysis*] (Yayımlanmamış doktora tezi). Ege Üniversitesi, İzmir.

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

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

Hong, S., Malik, M.L., & Lee, M. (2003). Testing configural, metric, scalar and latent mean invariance across genders in sociotropy and autonomy using a non- western sample. *Educational and Psychological Measurement*, *63*(4), 636-654. Retrieved from https://doi.org/10.1177/0013164403251332

Huang, C. D., Church, A. T., & Katigbak, M. S. (1997). Identifying cultural differences in items and traits: Differential item functioning in the NEO Personality Inventory. *Journal of Cross-Cultural Psychology*, *28*, 192-218. Retrieved from http://dx.doi.org/10.1177/0022022197282004

Kankaras, M., Vermunt, J. K., & Moors, G. (2011). Measurement equivalence of ordinal items: A comparison of factor analytic, item response theory, and latent class approaches. *Sociological Methods & Research*, *40*(2), 279-310. Retrieved from https://doi.org/10.1177/0049124111405301

Karasar, N. (2005). *Bilimsel araştırma yöntemi [Scientific research method].* Ankara: Nobel yayın dağıtım

Kazelskis, R., Thames, D., & Reeves, C. (2004). The elementary reading attitude survey: factor invariance across gender and race. *Reading Psychology*, *25*, 111-120. Retrieved from https://doi.org/10.1080/02702710490435682

Kim, S. H., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement*, *22*(4), 345-355. Retrieved from https://doi.org/10.1177/014662169802200403

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling*, *18*(2), 212-228. Retrieved from https://doi.org/10.1080/10705511.2011.557337

Korkmaz, M. (2005). *Madde cevap kuramına dayalı olarak çok kategorili maddelerde madde ve test yanlılığının (işlevsel farklılığın) incelenmesi. [An investigation of differential item and test functioning at polytomous items based on item response theory]*. (Doktora tezi) Ege Üniversitesi, İzmir.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*(1), 53-76. Retrieved from https://doi.org/10.1207/s15327906mbr3201_3

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*(4), 361-388. Retrieved from https://doi.org/10.1177/1094428104268027

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525-543. Retrieved from https://doi.org/10.1007/BF02294825

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, *18*(2), 5-11. https://doi.org/10.1002/j.2330-8516.1988.tb00303.x

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons’ responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741-749.

Morales, L. S., Reise, S. P., & Hays, R. D. (2000). Evaluating the equivalence of health care ratings by whites and Hispanics. *Medical care*, *38*(5), 517.

Muraki, E., & Bock, R. D. (1991). *PARSCALE: Parameter scaling of rating data.* Chicago: Scientific Software, Inc.

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology*, *5*, 978. https://doi.org/10.3389/fpsyg.2014.00978

Nachtigall, C., Kroehne, U., Funke, F., & Steyer, R. (2003). (Why) should we use SEM? Pros and Cons of Structural Equation Modeling. *Methods of Psychological Research Online*, *8*(2), 1- 22.

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

Reise, S. P., Smith, L., & Furr, R. M. (2001). Invariance on the NEO PI-R neuroticism scale. *Multivariate Behavioral Research*, *36*(1), 83-110. Retrieved from https://doi.org/10.1207/S15327906MBR3601_04

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*(3), 552-566. Retrieved from https://doi.org/10.1037/0033-2909.114.3.552

Rogers, J. H., & Swaminathan, H. (1993). A comparison of logistic regression and mantel-haenszel procedures for detecting differential ıtem functioning. *Applied Psychological Measurement*, *17*(2), 105-116. Retrieved from https://doi.org/10.1177/014662169301700201

Sırgancı, G., Uyumaz, G., & Yandı, A. (2020). Measurement Invariance Testing with Alignment Method: Many Groups Comparison. *International Journal of Assessment Tools in Education*, *7*(4), 657-673. https://doi.org/10.21449/ijate.714218

Somer, O. (2004). Gruplararası karşılaştırmalarda ölçek eşdeğerliğinin incelenmesi: Madde ve test fonksiyonlarının farklılaşması. *Türk Psikoloji Dergisi*, *19*(53), 69 - 82.

Somer, O., Korkmaz, M., Dural, S., & Can, S. (2009). Ölçme eşdeğerliğinin yapısal eşitlik modellemesi ve madde tepki kuramı kapsamında incelenmesi. [Detection of Measurement Equivalence by Structural Equation Modeling and Item Response Theory]. *Türk Psikoloji Dergisi*, *24*(64), 61-75.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*(6), 1292. Retrieved from https://doi.org/10.1037/0021-9010.91.6.1292

Tekin, H. (1991). *Eğitimde ölçme ve değerlendirme. [Measurement and evaluation in education]*. (6. baskı). Ankara: Yargı Yayınları.

Thissen, D. (1991). *MULTILOG: Multiple category item analysis and test scoring using item response theory*. Chicago, IL: Scientific Software International.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, *99*(1), 118. Retrieved from http://dx.doi.org/10.1037/0033-2909.99.1.118

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

Van de Vijver, F. J. (1998). Towards a theory bias and equivalence. *Zuma Nachrichten: Cross-Cultural Survey Equivalence*, *3*, 41-65. Retrieved from https://nbn-resolving.org/urn:nbn:de:0168-ssoar-49731-1

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4-69. Retrieved from https://doi.org/10.1177/109442810031002

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

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. *The science of prevention: Methodological advances from alcohol and substance abuse research*, 281-324. Retrieved from http://dx.doi.org/10.1037/10222-009

Wu, A. D., Li, Z., & Zumbo, B. D. (2007) Decoding the meaning of factorial invariance and updating the practice of multiple-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research and Evaluation*, 12, 1-26.

APPENDIX

Appendix 1. The statistical power ratios of MGCFA

Sample Size		Number of items	Ratios of the items having DIF	Statistical power ratios
Focal	Reference			
250	250	20	10	45.00
500	500	20	10	78.00
1000	1000	20	10	92.00
250	250	20	20	35.00
500	500	20	20	72.50
1000	1000	20	20	88.50
250	250	20	30	29.33
500	500	20	30	58.00
1000	1000	20	30	82.00
250	250	40	10	47.00
500	500	40	10	77.00
1000	1000	40	10	93.50
250	250	40	20	38.00
500	500	40	20	63.75
1000	1000	40	20	86.50
250	250	40	30	27.33
500	500	40	30	55.33
1000	1000	40	30	81.00
250	250	60	10	44.67
500	500	60	10	67.33
1000	1000	60	10	91.00
250	250	60	20	35.17
500	500	60	20	65.67
1000	1000	60	20	86.67
250	250	60	30	26.78
500	500	60	30	54.11
1000	1000	60	30	81.89

Appendix 2. The statistical power ratios of LRT.

Sample size		Number of items	Ratios of the items having DIF	Statistical power ratios
Focal	Reference			
250	250	20	10	59.75
500	500	20	10	79.75
1000	1000	20	10	91.25
250	250	20	20	61.00
500	500	20	20	76.33
1000	1000	20	20	88.83
250	250	20	30	60.33
500	500	20	30	75.25
1000	1000	20	30	87.00
250	250	40	10	81.00
500	500	40	10	93.00
1000	1000	40	10	96.75
250	250	40	20	78.25
500	500	40	20	90.75
1000	1000	40	20	95.15
250	250	40	30	75.50
500	500	40	30	88.25
1000	1000	40	30	93.33
250	250	60	10	39.00
500	500	60	10	56.25
1000	1000	60	10	69.50
250	250	60	20	37.00
500	500	60	20	55.75
1000	1000	60	20	66.25
250	250	60	30	36.25
500	500	60	30	52.50
1000	1000	60	30	62.75

Appendix 3. The statistical power ratios of CIP.

Sample size		Number of items	Ratios of the items having DIF	Statistical power ratios
Focal	Reference			
250	250	20	10	54.50
500	500	20	10	78.00
1000	1000	20	10	94.50
250	250	20	20	49.00
500	500	20	20	79.00
1000	1000	20	20	94.50
250	250	20	30	36.00
500	500	20	30	67.17
1000	1000	20	30	89.00
250	250	40	10	28.00
500	500	40	10	41.25
1000	1000	40	10	50.75
250	250	40	20	29.00
500	500	40	20	41.38
1000	1000	40	20	53.50
250	250	40	30	24.33
500	500	40	30	41.58
1000	1000	40	30	60.00
250	250	60	10	18.17
500	500	60	10	30.00
1000	1000	60	10	35.00
250	250	60	20	15.25
500	500	60	20	30.00
1000	1000	60	20	39.75
250	250	60	30	18.44
500	500	60	30	30.44
1000	1000	60	30	47.00

Appendix 4. Statistical power ratios of the methods for variables of sample size, number of items, and the ratios of the items having DIF

Sample size (F/R)	Number of items	The ratios of the items having DIF	MGCFA	LRT	CIP
250/250	20	10	45.00	59.75	54.50
500/500	20	10	78.00	79.75	78.00
1000/1000	20	10	92.00	91.25	94.50
250/250	20	20	35.00	61.00	49.00
500/500	20	20	72.50	76.33	79.00
1000/1000	20	20	88.50	88.83	94.50
250/250	20	30	29.33	60.33	36.00
500/500	20	30	58.00	75.25	67.17
1000/1000	20	30	82.00	87.00	89.00
250/250	40	10	47.00	81.00	28.00
500/500	40	10	77.00	93.00	41.25
1000/1000	40	10	93.50	96.75	50.75
250/250	40	20	38.00	78.25	29.00
500/500	40	20	63.75	90.75	41.38
1000/1000	40	20	86.50	95.15	53.50
250/250	40	30	27.33	75.50	24.33
500/500	40	30	55.33	88.25	41.58
1000/1000	40	30	81.00	93.33	60.00
250/250	60	10	44.67	39.00	18.17
500/500	60	10	67.33	56.25	30.00
1000/1000	60	10	91.00	69.50	35.00
250/250	60	20	35.17	37.00	15.25
500/500	60	20	65.67	55.75	30.00
1000/1000	60	20	86.67	66.25	39.75
250/250	60	30	26.78	36.25	18.44
500/500	60	30	54.11	52.50	30.44
1000/1000	60	30	81.89	62.75	47.00