DOI: 10.5455/msm.2018.30.121-126

Received: March 18 2018; Accepted: May 21, 2018

© 2018 Jamshid Jamali, Seyyed Mohammad Taghi Ayatollahi, Peyman Jafari

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

ORIGINAL PAPER

The Effect of Cross-loading on Measurement Equivalence of Psychometric Multidimensional Questionnaires in MIMIC Model: a Simulation Study

Jamshid Jamali^{1,2}, Seyyed Mohammad Taghi Ayatollahi³, Peyman Jafari³

ABSTRACT

Introduction: In recent years, Multiple Indicators Multiple Causes (MIMIC) model has been widely used to assess measurement invariance, called Differential Item Functioning (DIF) analyses, in psychological and medical studies. **Aim:** This simulation study aimed at assessing the effect of sample size, scale length, and magnitude of the uniform-DIF on detecting uniform-DIF with the MIMIC model when it has cross-loading in multidimensional scales. **Material and Methods:** In this Monte Carlo simulation study, we calculated power, Type I error rates, the bias of parameters estimation, Coverage Probability (CP), and Convergence Rate (CR) was used to assess the performance of the MIMIC model. The means of RMSEA, SRMR, CFI, and TLI, as indices of the goodness-of-fit for the MIMIC model, were computed across 1000 replications for each simulation condition. **Result:** Approximately, in all scenarios simulated, the bias of DIF parameters estimation was negligible. The existence of cross-loading caused a decrease of approximately 11.8% in the power and increase of 0.04-unit in bias parameter estimation. By increasing the relationship between dimensions, the power and CP of MIMIC model decreased, however, bias and CR were increased. In all scenarios that were performed in this study, all goodness-of-fit indices were at an acceptable level. **Conclusion:** Our results indicated that the performance of the MIMIC model improved, when sample size, the number of items, and the magnitude of DIF increased. When the scale is multidimensional and model have cross-loading, the performance of the MIMIC model becomes questionable.

Keywords: Reproducibility of Results, Monte Carlo Method, Psychometrics, Surveys and Questionnaires.

1. INTRODUCTION

Many scientific fields including medicine and psychometric studies deal with the design, construction or translation of questionnaires. The validity and reliability are the two most important concepts related to questionnaires. In recent years, measurement equivalence, also known as differential item functioning (DIF), has been widely used to evaluate the validity and reliability of questionnaires (1-5). DIF occurs when item's response function changes across different groups of respondents. It means that participants' perception of questionnaire's item differs across groups, after controlling for the certain construct (6, 7).

Numerous statistical approaches are designed to detect DIF (6, 8, 9). Structure Equation Modeling-based methods (SEM) with appropriate statistical properties are popular procedures for detecting DIF. Multiple Indicators Multiple Causes (MIMIC) model, an SEM-based DIF detection method, involves using latent variables that are predicted by observed variables (10). This method does not require unidimensionality and conditional independence assumptions (11). The MIMIC model does not need large sample size, provides information about the structural and measurement model, does not have a special restriction on the variable scale, each latent variable can be predicted by at least one observed indicator, and can

¹Department of Biostatistics and Epidemiology, School of Health, Mashhad University of Medical Sciences, Mashhad, Iran

²Social Determinants of Health Research Center, Mashhad University of Medical Sciences, Mashhad, Iran

³Department of Biostatistics, Faculty of Medicine, Shiraz University of Medical Sciences, Shiraz, IR Iran

Corresponding author: Seyyed Mohammad Taghi Ayatollahi. Department of Biostatistics, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran. Tel/Fax: +98-7132349330, E-mail: ayatolahim@sums.ac.ir ORCID ID: <https://orcid.org/0000-0002-5691-9628>

Figure 1. A MIMIC model used to detect uniform-DIF in this study

detect DIF in the multidimensional questionnaire (1, 12-16).

To identify DIF in the traditional MIMIC model, for each dimension of the instrument has to be fitted with one MIMIC model (17). Using this approach, apart from increasing the error rate, the relationship between instrument dimensions ignored. Consequently, in the recent year, several methods have been proposed to detect DIF in multidimensional questionnaires (18, 19). Divergent validity, also known as discriminant validity, is a technique to assess the construct validity of a questionnaire. In the well-designed questionnaire, an item must have a high correlation with the other items in the same dimension but have low correlation with items in other dimensions (20, 21). SEM-based applied research is typically including cross-loadings that it can lead to misspecifying the model (10). Misspecification of the MIMIC model can lead to biased coefficients and error terms, which in turn can produce incorrect inference in DIF testing (22-24). Up to now, cross-loading and the impact of relationships between instrument dimensions have not been examined in DIF literature. Therefore, this paper attempts to examine the performance of the MIMIC model in DIF detecting, while a cross-loading is present in the model, in addition, the relationships between the instrument dimensions were investigated.

2. MATERIALS AND METHODS

2.1. MIMIC Model for Detecting DIF

In DIF study, uniform (constant across ability levels) and non-uniform (varying across ability levels) can be identified,

but in our simulation study, we only looked at uniform-DIF. The uniform-DIF is important amongst applied researchers because the non-uniform-DIF has an interaction between ability level and group membership (17, 25). Uniform-DIF occurs when item thresholds differ between groups. In the MIMIC model, an item is tested for uniform-DIF by regressing it and latent variable (ability) onto a covariate simultaneously (26).

2.2. Simulation

Data and matrix generation are two methods for simulations in SEM. In DIF simulation studies using data generation, responses were generated from a common model, for example, Graded Response Model. Usually, if scale variables and relationships between the indicator variables with each other or with latent variables are important, the researcher uses the data generation. In order to generate raw data this method has higher accuracy and is more realistic. Due to the computational complexity of some matrix iterative algorithms in estimating MIMIC parameters, as well as the complex relationship between variable, hence we cannot always generate raw data. In matrix generation approach, the relationships between variables are redesigned with correlation and covariance matrix. Using this method leads to saving time, and provides the possibility of testing the hypotheses that cannot be evaluated by raw data. In this simulation study, five factors of sample size, the number of items in each dimension, the magnitude of uniform-DIF, the correlation between dimensions, and cross-loading were investigated. The sample size was set at 100, 250, and 500 for the small, moderate, and large sample

scale length		5-item scale								10-item scale							
Condition	magnitude of DIF	performance of the MIMIC model				Goodness of Fit				performance of the MIMIC model				Goodness of Fit			
		power	Bias	CP	CR	RMSEA	CFI	TLI	SRMR	power	Bias	CP	CR	RMSEA	CFI	TLI	SRMR
$\rho=0$ $r=0$	small	0.326	-0.025	0.969	76.9	0.045	0.969	0.959	0.044	0.430	0.01	0.972	88.6	0.032	0.972	0.969	0.048
	medium	0.748	-0.017	0.978	78.5	0.043	0.972	1.002	0.043	0.857	0.025	0.962	86.8	0.031	0.973	0.997	0.049
	large	0.898	0.015	0.968	78.3	0.044	0.971	0.962	0.043	0.960	0.015	0.950	87.5	0.031	0.975	0.972	0.049
$\rho=0$ $r=0.61$	small	0.145	0.021	0.931	84.1	0.044	0.973	0.963	0.043	0.194	-0.084	0.894	91.6	0.032	0.974	0.971	0.049
	medium	0.573	0.011	0.935	83.2	0.041	0.975	0.966	0.042	0.709	-0.038	0.875	89.6	0.032	0.974	0.971	0.048
	large	0.864	-0.011	0.928	79.2	0.044	0.974	0.966	0.042	0.939	-0.098	0.878	90.1	0.032	0.975	0.972	0.049
$\rho=0.25$ $r=0.61$	small	0.14	0.004	0.924	88.6	0.043	0.974	0.965	0.041	0.186	-0.087	0.865	94.4	0.031	0.975	0.972	0.047
	medium	0.582	-0.150	0.903	86.6	0.042	0.976	0.967	0.042	0.727	-0.086	0.854	95.2	0.032	0.975	0.972	0.047
	large	0.897	-0.093	0.902	87.6	0.044	0.974	0.965	0.042	0.954	-0.059	0.858	95.3	0.032	0.975	0.972	0.048
$\rho=0.5$ $r=0.61$	small	0.141	-0.077	0.885	92.5	0.042	0.978	0.97	0.04	0.170	-0.111	0.850	98.2	0.031	0.976	0.973	0.045
	medium	0.604	-0.010	0.882	91.5	0.042	0.976	0.968	0.041	0.745	-0.086	0.829	98.2	0.031	0.977	0.974	0.045
	large	0.93	-0.098	0.875	91.9	0.043	0.978	0.971	0.041	0.969	-0.094	0.840	98.2	0.031	0.976	0.973	0.046
$\rho=0.75$ $r=0.61$	small	0.126	-0.114	0.873	93.0	0.042	0.979	0.972	0.039	0.179	-0.112	0.826	98.2	0.03	0.979	0.976	0.042
	medium	0.646	-0.108	0.859	92.2	0.043	0.979	0.972	0.039	0.748	-0.116	0.813	98.9	0.031	0.979	0.976	0.043
	large	0.94	-0.146	0.859	93.4	0.041	0.981	0.974	0.039	0.983	-0.092	0.824	98.8	0.030	0.979	0.976	0.043

Table 1. Evaluation the impact of cross-loading on uniform DIF detection in MIMIC model in sample size 100. ρ : Correlation coefficient between the two dimensions; r : Cross-loading; MIMIC: Multiple Indicators Multiple Causes, DIF: Differential Item Functioning; CP: Coverage Probability; CR: Convergence Rate; RMSEA: Root Mean Squared Error of Approximate; SRMR: Standardized Root Mean Squared Residual; CFI: Comparative Fit Index; TLI: Tucker-Lewis Index

size, respectively. The number of items in each dimension was 5 and 10. The magnitude of DIF sizes, 0.25, 0.5 and 1 were used to illustrate small DIF, medium DIF, and large DIF sizes for the studied item. The four levels of correlation ($\rho=0$, 0.25, 0.5 and 0.75) were considered as no, poor, moderate and strong correlation between the two dimensions (27). Also, we created two conditions with and without cross-loading in the MIMIC model. In the present study, the regression coefficient showing the group mean difference on the factor for covariate was held constant at 0.5. In this study, all factor loadings for each dimension were 0.7. Some researchers have pointed out that if the difference between factor loadings and cross-loading is less than 0.1, attention should be given to the cross-loading since this model is misspecified (28). Hence, we considered the cross-loading amount equal to 0.61. Figure 1 illustrates a MIMIC model used to detect uniform-DIF in this study.

The power, bias of parameters estimation, coverage probability (CP), and convergence rate (CR) were used to assess the performance of the MIMIC model. Statistical power is defined by the proportion of items correctly identified as DIF averaged across replications (29). The bias of DIF parameter is the difference between the true and estimated values of the parameter. The CP was determined by the ratio of the number of times the 95% confidence interval for DIF item's parameter estimates contains the actual value parameters across 1000 replications. When any item correlates more strongly with the other dimensions than with its own dimension, we say that our instrument has cross-loading (21).

Appropriateness of the MIMIC model in DIF detection was assessed by four fit indices: Root Mean Squared Error of Approximate (RMSEA), Standardized Root Mean Squared Residual (SRMR), Comparative Fit Index (CFI), and Tucker-Lewis Index (TLI). In MIMIC model, judgments about how well the model fit the data were made based on RMSEA < 0.06, SRMR

< 0.08, CFI > 0.95, and TLI > 0.95 (30). In this Monte Carlo simulation study, each simulated scenario was replicated 1000 times, the nominal significance level was set to 0.05, and seed function was used to generate the same sets of random numbers and to reach result in each time run scenarios.

All simulations were performed with the *simsem* package in R. This package is an easy framework for Monte Carlo simulation in SEM, which can be used for various purposes, such as DIF study (31).

3. RESULTS

3.1. Sample size effect on detecting uniform-DIF in the MIMIC model

When other conditions stayed fixed, increasing the sample size led to improving the power of the MIMIC model. The mean±SD power of MIMIC model in the sample size of 100, 250 and 500 were 0.61±0.32, 0.72±0.27, and 0.77±0.21, respectively. The mean±SD of parameter estimation' bias in the sample size of 100, 250 and 500 were 0.07±0.05, 0.06±0.03, and 0.06±0.04, respectively, where a meaningful change was not observed. The mean±SD CP of DIF parameter in the sample size of 100, 250 and 500 were 0.89±0.05, 0.82±0.10, and 0.75±0.16, respectively, these numbers indicate that increase in sample size leads to reduce CP. The CR of iteration in the sample size of 100, 250 and 500 were 90.24±6.47, 86.94±8.81, and 83.80±10.27, respectively, which is an indication of reduced CR by increasing the sample size.

3.2. Number of item effect on detecting uniform-DIF in the MIMIC model

By considering that other factors were constant, increasing the number of items from 5 to 10, led to the improvement of 0.67±0.28 to 0.73±0.27 in the power of the MIMIC model. The mean±SD of bias, CP, and CR were 0.06±0.04, 0.85±0.12, and

scale length		5-item scale								10-item scale							
Condition	magnit- ude of DIF	performance of the MIMIC model				Goodness of Fit				performance of the MIMIC model				Goodness of Fit			
		power	Bias	CP	CR	RMSEA	CFI	TLI	SRMR	power	Bias	CP	CR	RMSEA	CFI	TLI	SRMR
$\rho=0$ $r=0$	small	0.534	0.060	0.961	71.6	0.028	0.988	0.984	0.027	0.665	0.023	0.945	80.5	0.019	0.991	0.990	0.030
	medium	0.828	0.057	0.979	72.5	0.029	0.988	1.001	0.026	0.890	0.017	0.964	80.9	0.018	0.991	1.000	0.030
	large	0.888	0.040	0.98	70.5	0.028	0.989	0.985	0.027	0.944	0.044	0.949	78.9	0.019	0.991	0.990	0.030
$\rho=0$ $r=0.61$	small	0.302	-0.020	0.902	78.4	0.028	0.989	0.985	0.027	0.342	-0.037	0.810	85.3	0.018	0.992	0.991	0.029
	medium	0.772	-0.009	0.911	77.6	0.028	0.989	0.985	0.026	0.849	-0.055	0.824	84.6	0.018	0.991	0.990	0.030
	large	0.898	0.022	0.915	74.5	0.028	0.989	0.986	0.026	0.949	-0.039	0.854	84.1	0.017	0.992	0.991	0.029
$\rho=0.25$ $r=0.61$	small	0.295	-0.082	0.84	84.0	0.027	0.99	0.986	0.026	0.380	-0.005	0.741	92.2	0.018	0.992	0.991	0.029
	medium	0.811	-0.059	0.834	82.7	0.027	0.99	0.986	0.026	0.881	-0.140	0.789	92.5	0.018	0.992	0.991	0.029
	large	0.918	-0.093	0.889	82.2	0.028	0.99	0.986	0.026	0.959	-0.066	0.780	92.8	0.018	0.992	0.991	0.029
$\rho=0.5$ $r=0.61$	small	0.25	-0.082	0.783	90.3	0.028	0.991	0.988	0.026	0.347	-0.081	0.682	95.6	0.018	0.992	0.991	0.028
	medium	0.831	-0.085	0.742	91.2	0.026	0.991	0.988	0.025	0.915	-0.097	0.737	96.6	0.017	0.993	0.992	0.028
	large	0.947	-0.093	0.795	92.1	0.028	0.991	0.988	0.026	0.974	-0.088	0.739	97.6	0.018	0.993	0.992	0.028
$\rho=0.75$ $r=0.61$	small	0.227	-0.053	0.722	94.0	0.026	0.992	0.989	0.024	0.337	-0.085	0.657	98.3	0.017	0.994	0.993	0.026
	medium	0.864	-0.076	0.722	95.1	0.026	0.992	0.99	0.024	0.927	-0.095	0.704	99.0	0.017	0.993	0.992	0.027
	large	0.952	-0.121	0.755	93.6	0.027	0.992	0.989	0.025	0.983	-0.094	0.737	99.0	0.017	0.994	0.993	0.026

Table 2. Evaluation the impact of cross-loading on uniform DIF detection in MIMIC model in sample size 250. ρ : Correlation coefficient between the two dimensions; r : Cross-loading; MIMIC: Multiple Indicators Multiple Causes, DIF: Differential Item Functioning; CP: Coverage Probability; CR: Convergence Rate; RMSEA: Root Mean Squared Error of Approximate; SRMR: Standardized Root Mean Squared Residual; CFI: Comparative Fit Index; TLI: Tucker-Lewis Index

scale length		5-item scale								10-item scale							
Condition	magnit- ude of DIF	performance of the MIMIC model				Goodness of Fit				performance of the MIMIC model				Goodness of Fit			
		power	Bias	CP	CR	RMSEA	CFI	TLI	SRMR	power	Bias	CP	CR	RMSEA	CFI	TLI	SRMR
$\rho=0$ $r=0$	small	0.588	0.072	0.973	67.8	0.020	0.994	0.992	0.018	0.682	0.014	0.935	73.8	0.013	0.996	0.995	0.020
	medium	0.823	0.053	0.975	67.2	0.018	0.995	1.001	0.018	0.898	0.036	0.943	75.4	0.013	0.996	1.000	0.020
	large	0.875	0.077	0.988	68.7	0.019	0.995	0.993	0.019	0.928	0.025	0.951	72.0	0.013	0.996	0.995	0.020
$\rho=0$ $r=0.61$	small	0.43	0.077	0.872	72.5	0.019	0.995	0.993	0.018	0.490	-0.074	0.700	82.2	0.013	0.996	0.995	0.020
	medium	0.803	-0.015	0.895	72.6	0.019	0.995	0.993	0.018	0.865	0.027	0.772	81.2	0.013	0.996	0.995	0.020
	large	0.886	0.017	0.909	71.7	0.019	0.995	0.993	0.018	0.924	0.036	0.834	77.6	0.013	0.996	0.995	0.020
$\rho=0.25$ $r=0.61$	small	0.444	0.100	0.798	78.9	0.019	0.995	0.993	0.018	0.495	-0.122	0.599	89.5	0.012	0.996	0.996	0.020
	medium	0.831	-0.033	0.797	79.7	0.019	0.995	0.993	0.018	0.900	-0.079	0.693	87.9	0.013	0.996	0.995	0.020
	large	0.927	-0.034	0.836	80.6	0.019	0.995	0.993	0.018	0.943	-0.061	0.722	87.1	0.012	0.996	0.996	0.020
$\rho=0.5$ $r=0.61$	small	0.446	0.026	0.685	87.4	0.018	0.996	0.994	0.018	0.494	-0.158	0.515	93.6	0.012	0.996	0.996	0.019
	medium	0.878	-0.020	0.672	88.2	0.018	0.996	0.994	0.018	0.919	-0.060	0.585	94.5	0.012	0.996	0.996	0.020
	large	0.955	-0.078	0.737	87.0	0.019	0.996	0.994	0.018	0.951	-0.044	0.637	95.7	0.012	0.996	0.996	0.020
$\rho=0.75$ $r=0.61$	small	0.380	-0.119	0.511	95.2	0.018	0.996	0.995	0.017	0.502	-0.100	0.488	97.8	0.012	0.997	0.996	0.018
	medium	0.906	-0.009	0.564	95.4	0.017	0.997	0.995	0.017	0.948	-0.121	0.556	97.9	0.012	0.997	0.996	0.019
	large	0.956	-0.065	0.612	95.8	0.017	0.997	0.995	0.017	0.980	-0.091	0.601	99.0	0.012	0.997	0.997	0.019

Table 3. Evaluation the impact of cross-loading on uniform DIF detection in MIMIC model in sample size 500. ρ : Correlation coefficient between the two dimensions; r : Cross-loading; MIMIC: Multiple Indicators Multiple Causes, DIF: Differential Item Functioning; CP: Coverage Probability; CR: Convergence Rate; RMSEA: Root Mean Squared Error of Approximate; SRMR: Standardized Root Mean Squared Residual; CFI: Comparative Fit Index; TLI: Tucker-Lewis Index

83.48±8.77 in 5-scale and 0.07±0.04, 0.79±0.13, and 90.50±7.76 in 10-scale, respectively. These results indicate that there was no trend in the bias, reduce the CP, and increase the CR by increasing number of items.

3.3. Magnitude of DIF effect on detecting uniform-DIF in the MIMIC model

When other factors are held constant during evaluation, with increasing severity of DIF, the power and CP of the MIMIC model were increased. However, no significant changes in bias and CR were observed. The mean±SD power in small,

medium and large DIF sizes were 0.36 ± 0.16 , 0.81 ± 0.10 , and 0.94 ± 0.03 , respectively. The mean $\pm$ SD bias in small, medium and large DIF sizes were 0.07 ± 0.04 , 0.06 ± 0.04 , and 0.06 ± 0.04 , respectively. The mean $\pm$ SD CP in small, medium and large DIF sizes were 0.80 ± 0.14 , 0.82 ± 0.12 , and 0.84 ± 0.11 , respectively. The mean $\pm$ SD CR in small, medium and large DIF sizes were 87.17 ± 8.78 , 87.11 ± 8.82 , and 86.70 ± 9.56 , respectively.

3.4. Relationship between dimensions' effect on detecting uniform-DIF in the MIMIC model

When there was no relationship between dimensions ($\rho=0$), the mean $\pm$ SD of power, bias, CP, and CR were 0.71 ± 0.24 , 0.04 ± 0.02 , 0.92 ± 0.07 , and 78.79 ± 6.71 , respectively. When the relationship between dimensions ($\rho=0.25$) was weak, the mean $\pm$ SD of power, bias, CP, and CR were 0.68 ± 0.29 , 0.08 ± 0.04 , 0.81 ± 0.08 , and 87.66 ± 5.39 , respectively. When the relationship between dimensions ($\rho=0.50$) was moderate, the mean $\pm$ SD of power, bias, CP, and CR were 0.69 ± 0.30 , 0.08 ± 0.04 , 0.75 ± 0.11 , and 93.35 ± 3.72 , respectively. When the relationship between dimensions ($\rho=0.75$) was strong, the mean $\pm$ SD of power, bias, CP, and CR were 0.70 ± 0.32 , 0.10 ± 0.03 , 0.70 ± 0.13 , and 96.37 ± 2.42 , respectively. The results of the simulation showed that when the relationship between dimensions increased, the power and CP were decreased and bias and CR were increased.

3.5. Presence of cross-loading effect on detecting uniform-DIF in the MIMIC model

When other conditions remained constant, the existence of cross-loading caused a decrease of approximately 11.8% in the power and increase 0.04-unit in bias parameter estimation. When the MIMIC model did not have a cross-loading factor, the mean $\pm$ SD of power, bias, CP, and CR were 0.76 ± 0.19 , 0.03 ± 0.02 , 0.96 ± 0.01 , and 76.47 ± 6.64 , respectively. When the MIMIC model had a cross-loading factor, the mean $\pm$ SD of power, bias, CP, and CR were 0.68 ± 0.29 , 0.07 ± 0.04 , 0.78 ± 0.11 , and 89.62 ± 7.41 , respectively.

In all scenarios that were performed in this study, all goodness-of-fit indices were at an acceptable level. For more information about goodness-of-fit indices and performance assessment indicators MIMIC model in uniform-DIF detection, refer to Tables 1 to 3.

4. DISCUSSION

Due to MIMIC model flexibility in DIF detection, this model has become an increasingly popular method to detect DIF. In this simulation-based study, we evaluated the impact of cross-loading and relationship between the dimensions on the performance of the MIMIC model in discovering uniform-DIF.

The results of our study showed that increasing the sample size, the number of items and severity of DIF could improve the performance of the MIMIC model for detecting uniform-DIF. In previous studies, similar results were obtained (2, 15, 16, 26, 32). The results of this study showed the relationship between the dimensions of the questionnaires that had led to reduce power and CP and increase the bias parameters. Lee et al., in their study found that increase relationship between the dimensions could lead to reduced power and increased the Type I error. This result concurs with the results of our study (19). A simulation study was conducted by Liaw revealed that

by increasing the relations between dimensions, multidimensional item response theory (MIRT) approach have a low bias in estimating the parameters and the relatively appropriate statistical power level in DIF detection (33). Finch et al., in their study, showed that by increasing the severity of the relationship between the dimensions of the questionnaire, bias, standard error, and root mean square error (RMSE) of item parameter estimates in two and three-parameter logistic MIRT was increased (34). In another study, Batley and Bose showed that standard deviation and standard error in the MIRT were not affected by the increased severity of the relationship between the dimensions of the questionnaire (35).

Another important feature considered in this study was the evaluation of cross-loading that could affect the performance of the MIMIC model. Cross-loading items are a major source of insufficient discriminant (36). Aside from the methodological problems, the results interpretation of MIMIC model with cross-loading is difficult. Previous studies have shown that the assessment of cross-loadings has not an appropriate measure for evaluating the discriminant validity in the questionnaire and this approach does not reliably identify the lack of discriminant validity in common research conditions (21). This method only works well in states with heterogeneous loading patterns and high sample sizes (21). However, the presence cross-loading in the model can lead to the misspecification the SEM (37). Our study showed that when the MIMIC model has a cross-loading factor, its performance is not appropriated. In other words, power and CP reduced and magnitude of the bias MIMIC model increased. However, the MIMIC model had acceptable power in uniform-DIF detection when the sample size was large or the magnitude of DIF was severe. To the best of our knowledge, the impact of cross-loading in DIF detection by the MIMIC model has not been investigated. Bauer in part of his study provided theoretical explanations about the impact of cross-loading on measurement equivalence in questionnaires (38). In this study, RMSEA, SRMR, CFI, and TLI indices were used to assess goodness-of-fit. As suggested by Taylor, RMSEA index is an appropriate index to assess goodness-of-fit in misspecification model (37). The results of our study indicated that all goodness-of-fit indices at acceptable levels; therefore, the results obtained in this study is reliable. Our study results should be viewed in light of its limitations. First, in this study, uniform-DIF was considered, although the MIMIC model could handle two type uniform and non-uniform-DIF. Second, we assume that the scale has two dimensions, one cross-loading, and one DIF items, while the MIMIC model can handle more conditions (19). Third, another limitation of this research was the relatively low number of reliable studies in this field to compare the result with it.

5. CONCLUSION

Our findings showed that the performance of the MIMIC model for uniform-DIF detection improved when sample size, the number of items, and the magnitude of DIF increased. When the scale is multidimensional and model have cross-loading, the performance of the MIMIC model for uniform-DIF detection are questionable. The results achieved from this study provide an appropriate guideline for further research in DIF study with the MIMIC model. We recommend further studies to investigate the effect of a number of items with DIF,

type of DIF, cross-loading, the relationship between dimensions of the performance of the MIMIC model for uniform-DIF detection when scale have more than two dimensions.

- **Abbreviation:** MIMIC: Multiple Indicators Multiple Causes DIF; Differential Item Functioning, CP: Coverage Probability, CR: Convergence Rate, RMSEA: Root Mean Squared Error of Approximate, SRMR: Standardized Root Mean Squared Residual, CFI: Comparative Fit Index, TLI: Tucker-Lewis Index, MIRT: Multidimensional Item Response Theory.
- **Declaration of Conflicting Interests:** The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.
- **Author's Contributions:** Jamshid Jamali contributed in planning, Statistical analysis and interpretation of results, writing, and critical revisions of the manuscript. Seyyed Mohammad Taghi Ayatollahi was responsible for the study conception and design, and supervising this study. Peyman Jafari contributed in planning, literature search, and manuscript editing.
- **Acknowledgments:** The authors would also like to thank Research Consultation Centre (RCC) for their invaluable editorial assistance.

REFERENCES

1. Jamali J, Ayatollahi SM, Jafari P. A New Measurement Equivalence Technique Based on Latent Class Regression as Compared with Multiple Indicators Multiple Causes. *Acta Inform Med*. 2016; 24(3): 168-171.
2. Jamali J, Ayatollahi SMT, Jafari P. The Effect of Small Sample Size on Measurement Equivalence of Psychometric Questionnaires in MIMIC Model: A Simulation Study. *Biomed Res Int*. 2017 ; Article ID 7596101.
3. Jafari P, Nozari F, Ahrari F, Bagheri Z. Measurement invariance of the Depression Anxiety Stress Scales-21 across medical student genders. *Int J Med Educ*. 2017; 8: 116-122.
4. Scott NW, Fayers PM, Aaronson NK, Bottomley A, de Graeff A, Groenvold M, et al. Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. *Health Qual Life Outcomes*. 2010; 8: 81.
5. Bagheri Z, Jafari P, Faghih M, Allahyari E, Dehesh T. Testing measurement equivalence of the SF-36 questionnaire across patients on hemodialysis and healthy people. *Int Urol Nephrol*. 2015; 47(12): 2013-2021.
6. Hidalgo MD, Gómez-Benito J. Differential Item Functioning. In: McGaw PPB, editor. *International Encyclopedia of Education* (Third Edition). Oxford: Elsevier; 2010: 36-44.
7. Teresi JA, Fleishman JA. Differential item functioning and health assessment. *Qual Life Res*. 2007; 16 Suppl 1: 33-42.
8. Doostfateme M, Taghi Ayatollah SM, Jafari P. Power and Sample Size Calculations in Clinical Trials with Patient-Reported Outcomes under Equal and Unequal Group Sizes Based on Graded Response Model: A Simulation Study. *Value Health*. 2016; 19(5): 639-647.
9. De Ayala RJ. *The Theory and Practice of Item Response Theory*. New York, USA: Guilford Press; 2013.
10. Schumacker RE, Lomax RG. *A beginner's guide to structural equation modeling*. 4 edition: Psychology Press; 2015.
11. Shih CL, Wang WC. Differential item functioning detection using the multiple indicators, multiple causes method with a pure short anchor. *Appl Psychol Meas*. 2009; 33(3): 184-199.
12. Ong ML, Lu L, Lee S, Cohen A, editors. *A comparison of the hierarchical generalized linear model, multiple-indicators multiple-causes, and the item response theory-likelihood ratio test for detecting differential item functioning*. Springer Proceedings in Mathematics and Statistics; 2015.
13. Muthén BO. Latent variable modeling in heterogeneous populations. *Psychometrika*. 1989; 54(4): 557-585.
14. Glockner-Rist A, Hoijtink H. The Best of Both Worlds: Factor Analysis of Dichotomous Data Using Item Response Theory and Structural Equation Modeling. *Struct Equ Modeling*. 2003; 10(4): 544-565.
15. Woods CM. Evaluation of MIMIC-Model Methods for DIF Testing With Comparison to Two-Group Analysis. *Multivariate Behav Res*. 2009; 44(1): 1-27.
16. Finch H. The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Appl Psychol Meas*. 2005; 29(4): 278-295.
17. Woods CM, Oltmanns TF, Turkheimer E. Illustration of MIMIC-Model DIF Testing with the Schedule for Nonadaptive and Adaptive Personality. *J Psychopathol Behav Assess*. 2009; 31(4): 320-330.
18. Cheng Y, Shao C, Lathrop QN. The Mediated MIMIC Model for Understanding the Underlying Mechanism of DIF. *Educ Psychol Meas*. 2016; 76(1): 43-63.
19. Lee S, Bulut O, Suh Y. Multidimensional Extension of Multiple Indicators Multiple Causes Models to Detect DIF. *Educ Psychol Meas*. 2017; 77(4): 545-569.
20. Hinkin TR. A Brief Tutorial on the Development of Measures for Use in Survey Questionnaires. *Organ Res Methods*. 1998; 1(1): 104-121.
21. Henseler J, Ringle CM, Sarstedt M. A new criterion for assessing discriminant validity in variance-based structural equation modeling. *J Acad Mark Sci*. 2015; 43(1): 115-135.
22. Swamy PAVB, Tavlaz George S, Hall Stephen GF, Hondroyannis G. Estimation of Parameters in the Presence of Model Misspecification and Measurement Error. *Studies in Nonlinear Dynamics and Econometrics*. 2010; 14(3): 1-33.
23. Chen F, Bollen KA, Paxton P, Curran PJ, KIRBY JB. Improper Solutions in Structural Equation Models. *Sociol Methods Res*. 2001; 29(4): 468-508.
24. Saris WE, Satorra A, van der Veld WM. Testing Structural Equation Models or Detection of Misspecifications? *Struct Equ Modeling*. 2009; 16(4): 561-582.
25. Woods CM, Grimm KJ. Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Appl Psychol Meas*. 2011; 35(5): 339-361.
26. Carroll IA. *MIMIC DIF Testing When the Latent Variable Variance Differs Between Groups*: University of Kansas, USA; 2014.
27. Walker CM, Sahin SG. Using a Multidimensional IRT Framework to Better Understand Differential Item Functioning (DIF): A Tale of Three DIF Detection Procedures. *Educ Psychol Meas*. 2016.
28. Hair JF, Hult GTM, Ringle CM, Sarstedt M. *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. Thousand Oaks, USA: Sage Publications; 2014.
29. Chun S. *Using MIMIC Methods to Detect and Identify Sources of DIF among Multiple Groups*. USA: University of South Florida; 2014.
30. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct Equ Modeling*. 1999; 6(1): 1-55.
31. Pornprasertmanit S, Miller P, Schoemann A. *SIMSEM: SIMulated Structural Equation Modeling*. 2016 [cited 2017 25 June]; Available from: <https://cran.r-project.org/web/packages/simsem/index.html>.
32. Willse JT, Goodman JT. Comparison of multiple-indicators, multiple-causes- and item response theory-based analyses of subgroup differences. *Educ Psychol Meas*. 2008; 68(4): 587-602.
33. Liaw YL. When can Multidimensional Item Response Theory (MIRT) Models be a Solution for Differential Item Functioning (DIF)? A Monte Carlo Simulation Study. Washington: University of Washington; 2015.
34. Finch H. Item Parameter Estimation for the MIRT Model. *Appl Psychol Meas*. 2010; 34(1): 10-26.
35. Batley RM, Boss MW. The Effects on Parameter Estimation of Correlated Dimensions and a Distribution-Restricted Trait in a Multidimensional Item Response Model. *Appl Psychol Meas*. 1993; 17(2): 131-141.
36. Farrell A, Rudd J, editors. *Factor analysis and discriminant validity: A brief review of some practical issues*. ANZMAC 2009 conference proceedings; 2009; Melbourne, Australia: ANZMAC.
37. Taylor AB. Two new methods of studying the performance of SEM fit indexes. Arizona, USA: Arizona State University; 2008.
38. Bauer DJ. A More General Model for Testing Measurement Invariance and Differential Item Functioning. *Psychol Methods*. 2017; 22(3): 507-526.