International Journal of Assessment Tools in Education
2022, Vol. 9, No. 3, 631–653
https://doi.org/10.21449/ijate.1076464
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Differential item functioning across gender with MIMIC modeling: PISA 2018 financial literacy items

Fatima Munevver Saaatcioglu[1,*]

[1]Ankara Yildirim Beyazit University, Rectorate, Ankara, Türkiye

ARTICLE HISTORY

Received: Feb. 20, 2022
Revised: June 27, 2022
Accepted: July 04, 2022

Keywords:
Differential item functioning,
Latent class analysis,
Measurement invariance,
Mixture modeling,
PISA 2018.

Abstract: The aim of this study is to investigate the presence of DIF over the gender variable with the latent class modeling approach. The data were collected from 953 students who participated in the PISA 2018 8th-grade financial literacy assessment in the USA. Latent Class Analysis (LCA) approach was used to identify the latent classes, and the data fit the three-class model better in line with fit indices. In order to obtain more information about the characteristics of the emerging classes, uniform and non-uniform DIF sources were identified by using the Multiple Indicator Multiple Causes (MIMIC) model. The findings are very important in terms of contributing to the interpretation of latent classes. According to the results, the gender variable was a source of DIF for latent classes. It is important to include direct effects by gathering unbiased estimates for the measurement and structural parameters. Disregarding these effects can lead to incorrect identification of implicit classess. A sample application of MIMIC model was performed in a latent class framework with a stepwise approach in this study.

1. INTRODUCTION

One of the basic aims of measurement studies is to develop and construct valid items measuring latent variable. In many studies, Differential Item Functioning (DIF) can be a threat to the validity of a test or a scale. DIF concept is defined that the situation in which “different groups of test takers with similar overall ability, or similar status on an appropriate criterion, have, on average, systematically different responses to a particular item” in AERA & NCME (2014). Definition of two types of DIF called uniform and non-uniform DIF was emphasized in the literature (Ackerman, 1992; Mellenbergh, 1982; Millsap & Everson, 1993; Swaminathan & Rogers, 1990). Uniform DIF occurs while students in one group consistently have a better chance of giving a correct answer than those with the similar ability level in another group. If the relationship is not consistent, in this situation non-uniform DIF occurs (Swaminathan & Rogers, 1990).

The concept of DIF is directly related to the concept of fairness and bias. Fairness means that for different groups of students, inferences made according to test scores are valid (ETS, 2019). Thus, fairness in the test is related to bias. If a fair test is applied, students with the similar level of competence have the similar probability of answering an item correctly. Therefore, items

*CONTACT: Fatima Munevver Saaatcioglu fmsaatcioglu@ybu.edu.tr Ankara Yildirim Beyazit University, Rectorate, Ankara, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

having DIF cause bias, which is a problem in psychological and educational testing. Bias is concerned to construct-irrelevant factors such as education, gender, culture, age, although students have the same trait or ability (Lee & Zhang, 2017; Messick, 1989). The construct is accepted as the test of interest and can explain the variance of student's performance in a test. According to Messick (1989), construct-irrelevant variance refers to variables unrelated to the construct, and it can occur when the test scores are affected by factors that are irrelevant to the construct. Test preparation, test development and administration, scoring, students' background knowledge, personality, answering strategies, and cognitive ability can be construct-irrelevant, and efforts are needed to minimize such effects (Gallagher et al., 2002; Haladyna & Downing, 2004). In addition, The Standards state that any bias causing students' scores in systematically high or low is construct-irrelevant variance (AERA & NCME, 2014).

Studies to examine item and test bias are based on two fundamental perspectives in measurement theory. First, from the Classic Test Theory (CTT) perspective, the Multi-group Confirmatory Factor Analysis (MG-CFA) method is tested for the relationship between observed variable and the latent trait for measurement invariance across groups. The second one is evaluated according to whether the ability levels of individuals in separate groups are equivalent on substance behavior with the Item Response Theory (IRT) and DIF approach (Embretson & Reise, 2000). In IRT framework for detecting DIF; differences in the probability to reply the item correctly for two groups are taken into account. Therefore, IRT methods focus on comparing item parameters of the groups i.e item characteristic curves (Thissen et al., 1993). In DIF studies that were conducted according to manifest grouping approach, assume that the groups come from a homogenous subgroups, and this homogenity means that items do not have DIF within the subgroups (De Ayala, et al., 2002). Latent classes can occur whether all students do not have homogeneous response patterns (De Ayala et al., 2002; Samuelsen, 2008). On the other hand, it is debated that DIF results obtained from groups may be biased (Rupp & Zumbo, 2006). Hence, it is proposed to use mixture models that reject the homogeneity of the data for DIF in latent classes. Mixture models consider a mixture of latent classes to compose the sample (Mislevy & Verhelst, 1990; Rost, 1990; De Ayala & Santiago, 2017). According to this mixture modeling approach, invariance assumption is no longer essential, and thus, item parameters are estimated for each latent class (Cho, 2007; Cohen & Bolt, 2005; De Mars & Lau, 2011; Oliveri, et al., 2013; Park et al., 2016; Rupp & Zumbo). Thanks to these studies, , DIF studies should be examined between latent classes. The MIMIC modeling is used by researchers within mixture modeling approach to explore the latent classes (Nylund-Gibson & Choi, 2018). In these studies, the researchers examined the effect of covariates on latent class variable. With this perspective, the direct effects can be examined from covariates to items determining possible sources of DIF, which is called MIMIC modeling (Masyn, 2017). Moreover, it can be examined if the identified latent classes are invariant whenthe students in a class have the similar responses (Kankaraš et al., 2011).

The MIMIC model can be defined as a form of Structural Equation Modeling (SEM). The model combines covariates into a CFA model. MIMIC model includes a measurement model enabling to detect the link within latent variable and items, and also a structural model bringing out the direct effect of a covariate. There are studies stating that MIMIC models are more useful compared to other techniques such as multigroup CFA in examining DIF (Vandenberg & Lance, 2000; Millsap, 2011). MIMIC modeling contributes to external validity by examining the relationship between covariate and latent structure, and to internal validity by estimating IRT parameters (Tsaousis et al., 2020).

MIMIC modeling allows us to see the effect of covariates. In addition, estimates can be obtained from all other grouping variables (covariates) in the model (Asparouhov & Muthén, 2015; Cheng et al., 2016). These variables can be observed or latent, and they can also be categorical or continuous (Glockner-Rist & Hoitjink, 2003). These flexibilities support the MIMIC model for DIF studies.

Next, in international large scale assessments like Program for International Student Assessment (PISA; OECD, 2019a) and Trends in International Mathematics and Science Study (TIMSS; IEA, 2017a) DIF analysis requires having the scores that are fairly comparable across countries. In international large scale assessments, IRT models are used to estimate item parameters. However, invariance assumption of the IRT models cannot be met in a heterogenous population which contains latent classes. Hence, the aim of this study is to investigate the presence of DIF over the gender variable with a stepwise procedure conducted with a MIMIC modeling framework that has been developed by Masyn, (2017). The MIMIC approach is a method to test measurement invariance, and since its introduction (Masyn, 2017), a study conducted with real data by Tsaousis et al., (2020) but there is no study with international large scale assessment data in which this method was used. Consequently, in this study, following the stepwise procedure outlined by Masyn (2017), to explore sources of DIF over gender using large scale assessment data (i.e., PISA 2018 financial literacy test). The results of this study are expected to have vital implications for measurement research by examining DIF between latent classes.

2. METHOD

2.1. Data

PISA is an international survey assessing competency of 15-year-old students in the basic domains of reading, mathematics and science literacy. PISA was first administered in 2000, and it cycles every three years. In each cycle, one of the basic domains is specified as the major domain, which is administered to all participants. The other domains are considered minor domains which are not administered to all participants. In addition, financial literacy was added in the PISA 2012 assessment, and has been provided as an international choice in the two PISA assessments (2015 and 2018). Financial literacy categories are money and transactions, planning and managing finance, risk and rewards, and an understanding of the financial landscape. These categories are measured by several open-ended and multiple-choice items (OECD, 2019a). In this study, 16 multiple choice items were used to detect uniform and non-uniform DIF items in Booklet 6. This booklet was used in the analysis because the number of items in the 6th booklet is more than the others in the booklets.

2.2. Sample

In PISA 2018**,** a total of 20 countries participated in financial literacy testing, including 13 OECD countries and seven partner countries. Since the purpose of this research is to show an application of the MIMIC model, the sample of this study includes 953 students from the USA who replied booklet 6. This country was chosen because the aim of this study is to show an application of the DIF study with MIMIC modeling and the sample size of the USA data is large. For the USA sample, 479 (50.3%) were females and 474 (49.7%) were males.

2.3. Data Analysis

The stepwise procedure has been developed by Masyn (2017), and in this study, the original source are used, and models are shown in Figure 1 with diyagrams for each step. First LCA was carried out to determine the number of latent classes. A procedure based on comparing the fit of models that have different numbers of latent classes and using model fit information criteria is applied in LCA. In simulation studies, it has been found that BIC outperforms in determining the number of latent classes (Nylund et al, 2007). In addition, sample-adjusted BIC is among the recommended indexes for consistent AIC (CAIC; Bozdogan, 1987) model fit. Next, VLMR and BLRT tests results are interpreted. LCA is performed with the Mplus (Muthén and Muthén, 1998-2021) software program using Robust Maximum-Likelihood (MLR) and expectation maximization (EM) algorithm as an estimation method.

This first step, i.e. Step 0 includes the process of deciding on the number of classes by finding the model that best fits the data with an exploratory approach. Considering the model fit indices, the number of latent classes are identified taking the covariate is as an auxiliary variable so that it does not have any effect on the determining latent classes (Masyn, 2013; Nylund-Gibson & Masyn, 2016). Only class indicators are included in the model as observed variables.

In Step 1, two models are compared. The first model, called A1.0, contains only the regression of the covariate over the latent class variable, which evaluates the model-fit of a non-DIF model. This model is compared to a model (non-uniform DIF) that items and latent variable are regressed to the covariate (A1.1) model in which the effects of the covariate on the items are released to differ between classes. The models compared with likelihood ratio test should supply proof on behalf of the A1.1 model as compared to A1.0 model, in the presence of DIF. If A1.0 model is the chosen model, there is no significant proof of DIF owing to the covariate. However, the choice of the A1.1 model requires further examination on the location of the invariance due to the covariate.

In Step 2, the purpose is to evaluate the presence of non-uniform DIF running models to detect the effects of the covariate on items. The models involve a no-DIF model (A2.0.1) that the latent class is regressing on the covariate and DIF model that an item regressing on the covariate A2.1.1 model from the first item to the last item. In model comparison, the likelihood ratio difference tests were utilized. Proof on behalf of the later model indicated the existence of non-uniform DIF.

In Step 3, the purpose is to select most parsimonious non-uniform DIF model (A3.0). This model helps to estimate a latent class model including non-uniform paths in which statistically significant. This model (A3.0) is first compared no-DIF model (A1.0) with the prospect that A3.0 would be excellent to model A1.0. The next comparison was between A3.0 and A1.1 (the all DIF model) with the expectancy that A3.0 would be no worse than A1.1.

In Step 4, we test hypothesis that non-uniform DIF effects items do not indicate uniform effects. Nonsignificant differences between models A3.0 and A4.1–A4.5, show proof of non-uniform DIF effects. Analyses were conducted with Mplus software (Muthén and Muthén, 1998-2019). The syntax codes for analyses can be found in the Appendix. In Figure 1, the model diagrams as stepwise procedure is given.

2.4. Effect Size

Several studies investigated the effect size metrics for DIF (Raju, 1990; Penfield & Lam, 2000; Zwick, 2012) and among them, the most considerable are the ETS criteria, transforming the difference logit parameter onto the delta metric system (Dorans & Holland, 1993). The Educational Testing Service (ETS) defined a three-level category sizes of DIF that are negligible, medium and large. For the negligible DIF level, the size of DIF should be 0.43 and below; for medium DIF, the size of the DIF should be ≥ 0.44 and for the large DIF, the size of the DIF should be ≥ 0.64 on logit scale (Lin & Lin, 2014).

Figure 1. *Stepwise procedure for DIF detection using mixture modeling.*

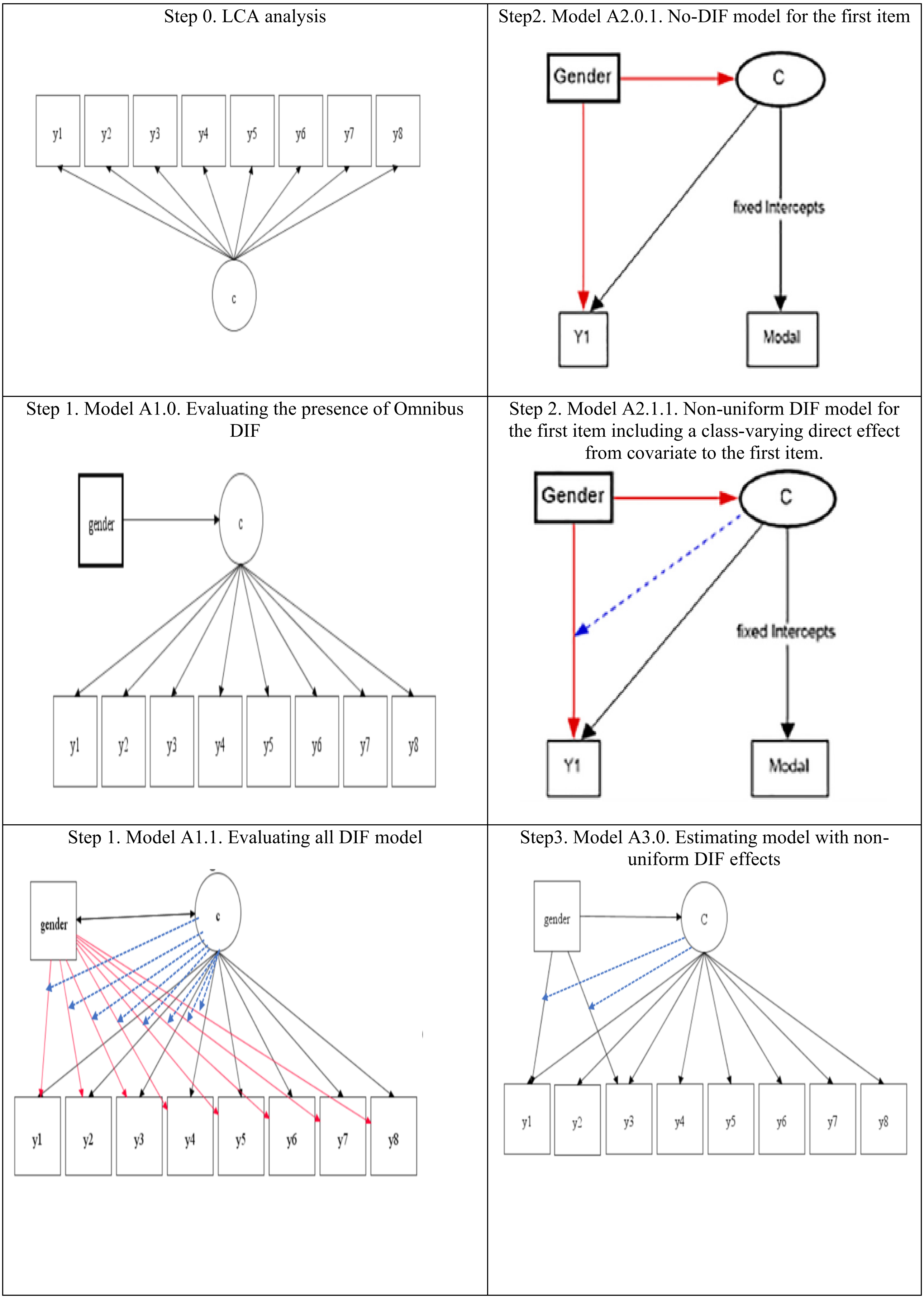
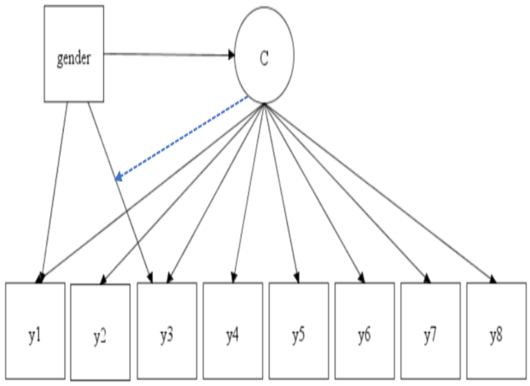


Figure 1. *Continues*

Step 4. Verifying DIF is non-uniform vs. uniform



3. RESULTS

3.1. Step 0

In this step, 1, 2, 3 and 4-class models were tested, respectively, and the model fit indices were presented in Table 1.

Table 1. *Fit indices of models tested for data from the USA.*

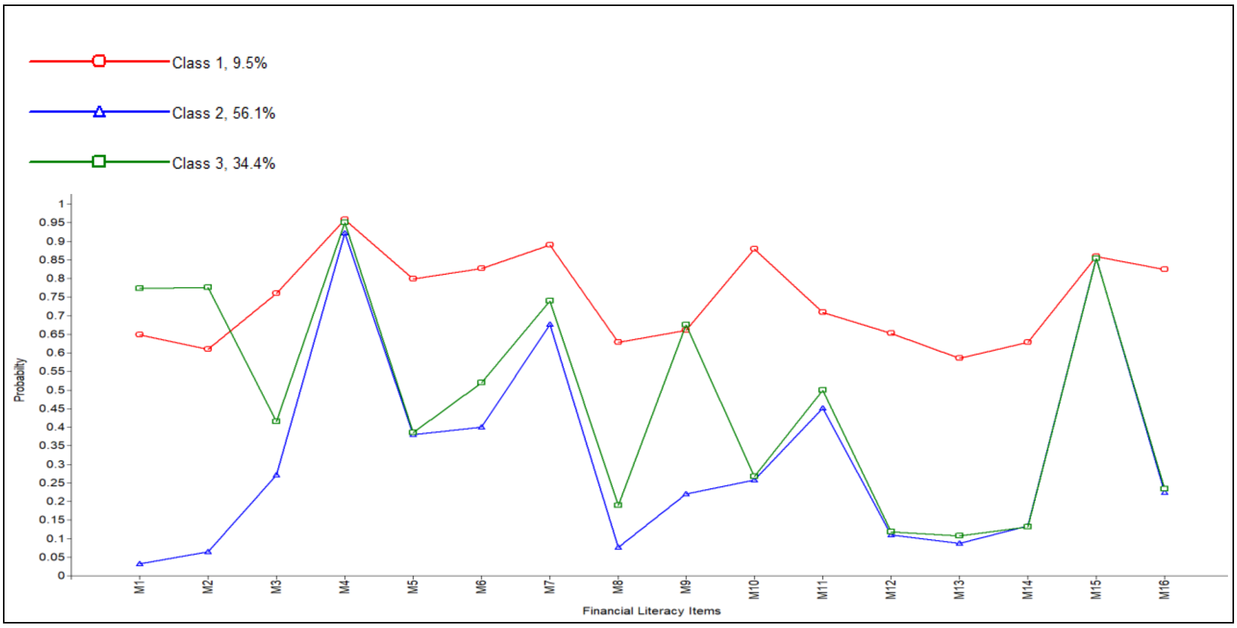
Fit indices	1-class	2-class	3-class	4-class
npar	16	33	50	67
LL	-8292.439	-7961.477	**-7727.144**	-7814.792
AIC	16616.878	15988.954	**15588.287**	15729.585
BIC	16694.598	16149.251	**15913.741**	15972.460
SSA-BIC	16643.783	16044.445	**15700.952**	15813.663
LR Chi-Square Test	2825.081	2463.731	2440.565	2275.961
LR Chi-Square *p* value	1.0000	1.0000	1.0000	1.0000
VLMR Test	-	656.295	293.369	175.297
VLMR *p* value	-	0.0000	0.0000	0.0649
BLRT Test	-	656.295	293.369	175.297
BLRT *p* value	-	0.0000	0.0000	0.0000

npar, number of free parameters; LL, log likelihood; AIC, Akaike's information criterion; BIC, Bayesian information criterion; SA-BIC, Sample-Size Adjusted BIC; VLMR, Vuong-lo-mendell-rubin test; BLRT, Bootstrapped likelihood ratio test; $p < 0.05$.

When the fit indexes were examined in the Table 1, the LR Chi-Square test, has an insignificant *p* value for data from the USA, showing that the model data fit was achieved. When the AIC, BIC and SSA-BIC values of the relative fit indices were examined, the three-class model had the lowest values among other models. The results of VLMR and BLRT showed a statistically significant difference between the 2-class and 3-class models. When the 3-class model was compared with the 4-class model, it was observed that the *p* value was not significant ($p>0.05$). This finding means that to add one more class to the 3-class model does not improve the model-data fit. As a result, it was seen that a three-class model was fit to the data. The model classified 9.5% students into Class 1 which had high probability of ability, 56.1 % subjects into Class 2 which had moderate probability of ability and 34.4% in Class 3 with low probability of ability.

The value of classification accuracy was 0.78. It can be stated that the three-class model is useful in assigning students to the correct classes as the entropy value was obtained to be greater than .70 (Nagin, 2005). The graph of the results obtained for these three classes were presented in Figure 2.

Figure 2. *Latent class profile plots for PISA 2018 financial literacy subtest.*



After determining the best-fitting model, it was examined for DIF effects on the financial literacy items, MIMIC modeling results were given step by step.

3.2. Step 1

In this step, Null model (A1.0) assuming no DIF and an alternate model (A1.1) assuming DIF for all items were compared. The results from the likelihood ratio test statistics (LRTS) indicated that gender was a source of DIF rejecting the null model A1.0 (LRTS = 130.40, *df* = 48, *p* = 0.0001). Therefore, the analyses in Step 2 were performed to explore the item-level effects of the DIF source based on the result from the omnibus DIF finding.

3.3. Step 2

In this step, null model (A2.0.1- no DIF model) and an alternate model (A2.1.1 non-uniform DIF model) for a specific item were compared. The results obtained from LRTS were presented in Table 2. According to results for five items (9, 12, 13, 15, 16), the no DIF model was rejected on behalf of the alternate model. These results mean that the non-uniform DIF items can be differentiated over gender.

3.4. Step 3

In this step, A3.0 model constructed from items displaying non-uniform DIF from step 2. A3.0 model was compared to model A1.0 (no-DIF), and the latter showed fit (LRTS = 56.634, *df* = 15, *p* = 0.0001). When A3.0 model was compared to A1.1 model (all DIF), significant differences were found (LRTS = 73.766, *df* = 33, *p* = 0.0001). In addition, the BIC values for A3.0 model, it was 16054.657, and for A1.1 model, it was 16207.187. Thus, the model A3.0 that has lower BIC value was the preferred model compared to A1.1 model. The results were presented in Step 3 part of Table 2. Finally, results from this step recommended that A3.0 model was the last latent class MIMIC model.

Table 2. *Model comparisons for DIF by stepwise procedure.*

Steps	Model	Model description	LogL	npar	Model Comparison	LRTS	df	*p*-value
1	A1.0	MIMIC: No DIF	-7815.632	55	A1.0 vs. A1.1	130.40	48	0.0001
	A1.1	MIMIC: All DIF	-7750.432	103				
2	A2.0.1	Item1: No DIF	-1128.703	7	A2.0.1 vs. A2.1.1	3.408	3	Ns
	A2.1.1	Item1: NON-U DIF	-1126.99	10				
	A2.0.2	Item2: No DIF	-1202.695	7	A2.0.2 vs. A2.1.2	6.872	3	Ns
	A2.1.2	Item2: NON-U DIF	- -1199.259	10				
	A3.0.3	Item3: No DIF	-1434.869	7	A2.0.3 vs. A2.1.3	2.94	3	Ns
	A3.1.3	Item3: NON-U DIF	-1434.722	10				
	A4.0.4	Item4: No DIF	-1081.621	7	A2.0.4 vs. A2.1.4	1.910	3	Ns
	A4.1.4.	Item4: NON-U DIF	-1080.666	10				
	A5.0.5.	Item5: No DIF	-1459.193	7	A2.0.5 vs. A2.1.5	1.644	3	Ns
	A5.1.1	Item5: NON-U DIF	-1458.371	10				
	A6.0.6	Item6: No DIF	-1476.005	7	A2.0.6 vs. A2.1.6	0.848	3	Ns
	A6.1.6	Item6: NON-U DIF	-1475.581	10				
	A7.0.7	Item7: No DIF	-1405.517	7	A2.0.7 vs. A2.1.7	1.216	3	Ns
	A7.1.7	Item7: NON-U DIF	-1404.909	10				
	A8.0.8	Item8: No DIF	-1205.034	7	A2.0.8 vs. A2.1.8	0.892	3	Ns
	A8.1.8	Item8: NON-U DIF	-1204.588	10				
	A9.0.9	Item9: No DIF	-1382.867	7	A2.0.9 vs. A2.1.9	10.658	3	0.01
	A9.1.9	Item9: NON-U DIF	-1377.538	10				
	A10.0.10	Item10: No DIF	-1379.947	7	A2.0.10 vs. A2.1.10	3.460	3	Ns
	A10.1.10	Item10: NON-U DIF	-1378.217	10				
	A11.0.11	Item11: No DIF	-1486.230	7	A2.0.11 vs. A2.1.11	5.94	3	Ns
	A11.1.11	Item11: NON-U DIF	1483.998	10				
	A12.0.12	Item12: No DIF	-1202.668	7	A2.0.12 vs. A2.1.12	10.792	3	0.01
	A12.1.12	Item12: NON-U DIF	-1197.272	10				
	A13.0.13	Item13: No DIF	-1176.637	7	A2.0.13 vs. A2.1.13	10.180	3	0.01
	A13.1.13	Item13: NON-U DIF	-1171.547	10				
	A14.0.14	Item14: No DIF	-1242.057	7	A2.0.14 vs. A2.1.14	1.142	3	Ns
	A14.1.14	Item14: NON-U DIF	-1241.486	10				
	A15.0.15	Item15: No DIF	-1238.851	7	A2.0.15 vs. A2.1.15	9.382	3	0.02
	A15.1.15	Item15: NON-U DIF	-1234.160	10				
	A16.0.16	Item16: No DIF	-1343.979		A2.0.16 vs. A2.1.16	8.698		0.03
	A16.1.16	Item16: NON-U DIF	-1339.630					
3	A3.0	MIMIC with all NON-U DIF items	-7787.315	70	A1.0 vs. A3.0	56.634	15	0.0001
					A3.0 vs. A1.1	73.766	33	0.0001
4	A4.1	Item9 (U- DIF) all other (NON-U DIF)	-7791.601	67	A4.1 vs. A3.0	8.572	3	0.035
	A4.2	Item12 (U-DIF) all other (NON-U DIF)	-7796.839	67	A4.2 vs. A3.0	19.048	3	0.0001
	A4.3	Item13 (U- DIF) all other (NON-U DIF)	-7795.662	67	A4.3 vs. A3.0	16.694	3	0.0001
	A4.4	Item15 (U- DIF) all other (NON-U DIF)	-7792.602	67	A4.4 vs. A3.0	10.574	3	0.014
	A4.5	Item16 (U -DIF) all other (NON-U DIF)	-7792.077	67	A4.5 vs. A3.0	9.524	3	0.023

LL: log likelihood; *df*.: degrees of freedom; LRTS: likelihood ratio test statistic, ;npar: number of free parameters, UN-DIF: uniform DIF, NON-U DIF: Non-uniform DIF, Ns: not significant; $p < 0.05$.

3.5. Step 4

In this step, MIMIC models (A4.1-A4.5) including the items which displayed non-uniform DIF. In these models, all other direct effects were allowed to free all across classes but the direct effect to each item was constrained to be invariant. Hence, each model (A4.1-A4.5) was compared with the non-uniform DIF model (A3.0 model). According to the results, it was found that models were statistically worse than A3.0 model, and DIF effects were non-uniform DIF (items 9, 12, 13, 15, 16).

Table 3. *Statistics for non-uniform DIF items over gender for PISA 2018 financial literacy subtest.*

	Latent Class 1			Latent Class 2			Latent Class 3		
Item no	Estimates	95% CIs (UL/LL)	Effect size	Estimates	95% CIs (UL/LL)	Effect size	Estimates	95% CIs (UL/LL)	Effect size
9	1.315	0.927/ 14.965	Large	-0.298	0.451/ 1.222	Negligible	-0.433	0.371/ 1.135	Medium
12	1.657	0.397/ 69.242	Large	1.080	1.428/ 6.076	Large	0.179	0.541/ 2.648	Negligible
13	0.978	0.496/ 14.269	Large	1.326	1.367/ 10.377	Large	0.206	0.523/ 2.884	Negligible
15	-1.216	0.011/ 7.669	Large	-0.598	0.321/ 0.943	Medium	-0.565	0.283/ 1.141	Medium
16	0.618	0.231/ 14.910	Medium	0.268	0.826 / 2.070	Negligible	0.748	1.156/ 3.860	Large

UL: upper level; LL: low level.

Table 3 presents estimates in logits, 95% CIs and effect size values for each classes. According to ETS criteria, the size of DIF effects were interpreted (Lin & Lin, 2014). For Class 1(high performing) item 9, 12, 13 and 15 exhibited large level DIF, and item 16 showed medium level DIF over gender. Moreover, males scored higher than females (positive values mean that males have higher values) on all items except item 15. For Class 2 (average performing), the DIF effect was negligible for item 9 and 16; item 12 and 13 showed large level DIF, and item 16 showed medium level DIF over gender. Also males scored higher than females on item 12 and 13. For Class 3 (low performing), the DIF effect was negligible for item 12 and 13, and it was medium for item 9 and 15 with males scoring higher than females; and it was large for item 16 with females scoring higher than males.

4. DISCUSSION and CONCLUSION

The aim of this study was to investigate the presence of DIF over the gender variable with a MIMIC modeling including a stepwise procedure (Masyn, 2017). In the first step, a LCA was conducted to detect group of heterogeneity. According to the indices, data fit the three-class model better. The model classified 9.5% of the students into Class 1(high performing), 56.1 % of the students into Class 2 (average performing) and 34.4% in Class 3 (low performing).

In addition to the above classification of the students into the three classes, this analysis could provide further information about the specific items that performed across the different classes. For example, item 4 was an easy item and had a high probability of ability for each class. A similar pattern was observed with item 7 and 15. Also item 5 and 10 seem to be difficult items that discriminate Class 1 (high performing) from the Class 2 (average performing) and Class 3 (low performing) but not differentiate Class 2 and Class 3 (average and low performing). It can be seen that the majority of the items differentiate students across classes.

Then, it was investigated if there had been direct effects from the latent class to items. Thus, DIF test was conducted by comparing no DIF model with all-DIF model considering no DIF model was statistically worse than all DIF model. So it can be stated that gender is a source of DIF. This is an important result showing that gender should be added in the regression model. Studies reveal that ignoring the effects of covariates may lead to misspecifications for the latent classes (Asparouhov & Muthén, 2014; Clark & Muthén, 2009; Masyn, 2017).

Next, uniform and non-uniform DIF effects were investigated for financial literacy items. According to the results, five items displayed non-uniform DIF with significant *p* values. Next, the effect size of non-uniform DIF items was investigated over gender. For Class1, item 9, 12, 13 and 15 exhibited large level DIF effect, and item 16 exhibited medium level DIF effect, when males scored higher. For Class 2, the DIF effect was negligible for item 9 and 16; item 12 and 13 exhibited large level DIF effect, and item 16 exhibited medium level DIF effect over gender. Furthermore, males also scored higher than females on item 12 and 13. For Class 3, the DIF effect was negligible for item 12 and 13, medium for item 9 and 15 with males who had higher scores; large for item 16 with females who had higher scores.

This study showed that what may be the cause for DIF in a latent class framework. Ignoring DIF effects in LCA could lead to the misinterpretation of the analysis and getting biased estimates in identifying classes and estimating relationship between latent class variable and covariate. Previous studies have shown that ignoring these effects can lead to biased estimated parameters for both measurement and structural model of the latent class analysis, and in this situation latent classes cannot be used for class comparisons (Clark & Muthén, 2009; Masyn, 2017; Nylund-Gibson & Choi, 2018). The results showed that MIMIC modeling was an essential procedure to find items displaying DIF effects between females and males. Thus, the nature of latent classes may be investigated by considering in each latent class membership. In addition in LCA, direct effects examinations must be a standard procedure to investigate direct effects of covariates on latent class indicators.

This study also revealed that response probabilities across latent classes were not the same for all latent class indicators. In this context, students within a class could have different response probabilities depending on a specific characteristic (in terms of gender). Hence, it can be pointed out that assuming that all latent class indicators have the same expected responses across classes and across different levels of a demographic variable can lead to the misinterpretation of latent classes. Thus, identifying latent classes by inspecting the manifest characteristics in each latent class membership is so important to have the right information about classes.

Throughout this article, the analyses were conducted on logit scale in Mplus, and the effect sizes were interpreted according to logit scale. The MIMIC model can be conducted for logistic or normal-ogive link functions. Thus, analysis can be run on probit link.

MIMIC modeling contributes to external validity by examining the relationship between covariate and latent structure, and to internal validity by estimating the parameters. Contributing to validity studies, other demographic variables can be included in the analysis. Next, various distal outcomes could be used to detect latent classes displaying statistically significant differences.

Continuous or categorical variables and the mixture of both can be used in MIMIC model approach. In this study, dichotomous variables were used. Future studies can be conducted with continuous variables, and the models can be compared with information criteria like SRMR, TLI, CFI etc. model fit statistics (Kang & Cohen, 2007).

Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author.

Orcid

F. Munevver Saatcioglu https://orcid.org/0000-0003-4797-207X

REFERENCES

Ackerman, T.A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91. https://doi.org/10.1111/j.1745-3984.1992.tb00368.x

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345-370. https://doi.org/10.1007/BF02294361

Cheng, Y., Shao, C., & Lathrop, Q.N. (2016). The mediated MIMIC model for understanding the underlying mechanism of DIF. *Educational and Psychological Measurement*, *76*(1), 43-63. https://doi.org/10.1177/0013164415576187

Cho, S.J. (2007). *A multilevel mixture IRT model for DIF analysis* [Unpublished doctoral dissertation]. University of Georgia: Athens.

Choi, Y., Alexeev, N., & Cohen, A.S. (2015). Differential item functioning analysis using a mixture 3-parameter logistic model with a covariate on the TIMSS 2007 mathematics test. *International Journal of Testing, 15(3),* 239-253. https://doi.org/10.1080/15305058.2015.1007241

Clark, S.L., & Muthén, B. (2009). Relating latent class analysis results to variables not included in the analysis. Avaliable online at: http://www.statmodel.com/download/relatinglca.pdf

Cohen, A.S., & Bolt, D.M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42(2),* 133–148. https://doi.org/10.1111/j.1745-3984.2005.00007

De Ayala, R.J., Kim, S.H., Stapleton, L.M., & Dayton, C.M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing*, *2*(3-4), 243-276. https://doi.org/10.1080/15305058.2002.9669495

De Ayala, R.J., & Santiago, S.Y. (2017). An introduction to mixture item response theory models. *Journal of School Psychology*, *60*(1), 25-40. https://doi.org/10.1016/j.jsp.2016.01.002

Educational Testing Service. (2019). Standards for Quality and Fairness. Retrieved from https://www.ets.org/s/about/pdf/standards.pdf.

Embretson, S.E., & Reise, S.P. (2000). *Item Response Theory for Psychologists*. Erlbaum.

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

Finch, W.H., & French, B.F. (2012). Parameter estimation with mixture item response theory models: A Monte Carlo comparison of maximum likelihood and Bayesian methods. *Journal of Modern Applied Statistical Methods*, *11*(1), 14.

Gallagher, A., Bennett, R.E., Cahalan, C., & Rock, D.A. (2002). Validity and fairness in technology- based assessment: detecting construct- irrelevant variance in an open-ended, computerized mathematics task. *Educational Assessment, 8(1)*, 27-41. https://doi.org/10.1207/S15326977EA0801_02

Glockner-Rist, A., & Hoitjink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling, 10(4),* 544-565. https://doi.org/10.1207/S15328007SEM1004_4

Haladyna, T.M., & Downing, S.M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Assessment Issues and Practice, 23 (1),* 17-27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

IEA. (2017a). *TIMSS 2015 user guide for the international database*. Chestnut, MA: Lynch School of Education, Boston College & International Association for the Evaluation of Educational Achievement (IEA).

Kang, T., & Cohen, A.S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement, 31*(1), 331-358. https://doi.org/10.1177%2F0146621606292213

Kankaraš, M., Moors, G., & Vermunt, J.K. (2011). Testing for measurement invariance with latent class analysis. In E. Davidov, P. Schmidt, J. Billiet, & B. Mueleman (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 359–384). Routledge.

Lee, Y., & Zhang, J. (2017). Effects of differential item functioning on examinees' test performance and reliability of test. *International Journal of Testing*, *17(1),* 23–54. https://doi.org/10.1080/15305058.2016.1224888

Lin, P.-Y., and Lin, Y.-C. (2014). Examining student factors in sources of setting accommodation DIF. *Educational and Psychological Measurement 74*(1), 759–794. https://doi.org/10.1177%2F0013164413514053

Masyn, K. (2013). "Latent class analysis and finite mixture modeling," in *The Oxford handbook of quantitative methods in psychology*, Vol. 2, ed. T. D. Little (Oxford University Press), 551–611.

Masyn, K. (2017). Measurement invariance and differential item functioning in latent class analysis with stepwise multiple indicator multiple cause modeling. *Structural Equation Modeling: A Multidisciplinar Jounal, 24(2)*, 180-197. https://doi.org/10.1080/10705511.2016.1254049

Messick, S. (1989). "Validity," in *Educational Measurement*. Editor R. L. Linn 3rd ed. (NewYork: American Councilon Education and Macmillan), 13–103.

Millsap, R.E. (2011). *Statistical approaches to measurement invariance*. Taylor & Francis.

Mislevy, R.J., & Verhelst, N.D. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*(1), 195-215.

Nylund, K.L., Asparouhov, T., & Muthén, B.O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural equation modeling: A multidisciplinary Journal, 14(4)*, 535-569. https://doi.org/10.1080/10705510701575396

Nagin, D. (2005). Group-based modeling of development. Harvard University Press.

Nylund-Gibson, K., & Masyn, K.E. (2016). Covariates and mixture modeling: results of a simulation study exploring the impact of misspecified effects on class enumeration. *Structural Equation Modeling, 23*(1), 782-797. https://doi.org/10.1080/10705511.2016.1221313

OECD. (2019a). *PISA 2018 assessment and analytical framework*. PISA, OECD Publishing.

OECD. (2019b). *Technical report of the Survey of Adult Skills (PIAAC) (3rd Edition)*. OECD Publishing.

Oliveri, M.E., & von Davier, M. (2017). Analyzing the invariance of item parameters used to estimate trends in international large-scale assessments. In H. Jiao & R. W. Lissitz (Eds.), *Test fairness in the new generation of large-scale assessment* (pp. 121–146). Information Age Publishing, Inc.

Oliveri, M., Ercikan, K., & Zumbo, B. (2013). Analysis of sources of latent class differential item functioning in international assessments. *International Journal of Testing*, *13(3),* 272–293. https://doi.org/10.1080/15305058.2012.738266

Oliveri, M.E., & Ercikan, K. (2011). Do different approaches to examining construct comparability in multilanguage assessments lead to similar conclusions? *Applied Measurement in Education, 24(4)*, 349-366. https://doi.org/10.1080/08957347.2011.607063

Penfield, R.D., & Lam, T.C.M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19*(1), 5–15. https://doi.org/10.1111/j.1745-3992.2000.tb00033.x

Raju, N. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*(1), 197–207. https://doi.org/10.1177/014662169001400208

Rost, J. (1990). Rasch Models in Latent Classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271–282.

Samuelsen, K.M. (2008). Examining differential item functioning from a latent mixture perspective. In Hancock, G.R., & Samuelsen, K.M. (Eds.) *Advances in latent variable mixture models*, Information Age.

Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27(4)*, 361–370.

Tsaousis, I., Sideridis, G.D., & AlGhamdi, H.M. (2020). Measurement invariance and differential ıtem functioning across gender within a latent class analysis framework: evidence from a high-stakes test for university admission in Saudi Arabia. *Frontiers in Psychology, 11*, 1-13. https://doi.org/10.3389/fpsyg.2020.00622

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(1),* 4-69. https://doi.org/10.1177%2F109442810031002

Yun, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* [Unpublished doctoral dissertation]. University of California, Los Angeles.

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series, 2012(1)*, i-30. http://dx.doi.org/10.1002/j.2333-8504.2012.tb02290.x

APPENDIX

A.0

```
TITLE: Stepwise MIMIC Model DIF Detection
DATA: file is b16-gender.dat;
VARIABLE:
NAMES ARE gender m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16;
MISSING ARE ALL (99);
Auxiliary = gender;
USEVARIABLES ARE m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15
m16;
CATEGORICAL ARE m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16;
CLASSES = c(3);
ANALYSIS:
type=mixture;

MODEL:
%OVERALL%
[ c#1*-1.291 ];
[ c#2* 0.489 ];

c on gender;
  %C#1%
[ m1$1*  -0.612  ];
[ m2$1*  -0.444  ];
[ m3$1*  -1.149  ];
[ m4$1*  -3.141  ];
[ m5$1*  -1.383  ];
[ m6$1*  -1.561  ];
[ m7$1*  -2.091  ];
[ m8$1*  -0.524  ];
[ m9$1*  -0.668  ];
[ m10$1* -1.982  ];
[ m11$1* -0.891  ];
[ m12$1* -0.631  ];
[ m13$1* -0.345  ];
[ m14$1* -0.524  ];
[ m15$1* -1.815  ];
[ m16$1* -1.545  ];

%C#2%
[ m1$1*   3.386  ];
[ m2$1*   2.674  ];
[ m3$1*   0.990  ];
[ m4$1*  -2.473  ];
[ m5$1*   0.490  ];
[ m6$1*   0.404  ];
[ m7$1*  -0.740  ];
[ m8$1*   2.497  ];
[ m9$1*   1.261  ];
[ m10$1*  1.058  ];
```

```
[ m11$1*  0.194   ];
[ m12$1*  2.084   ];
[ m13$1*  2.346   ];
[ m14$1*  1.864   ];
[ m15$1* -1.777   ];
[ m16$1*  1.240   ];

%C#3%
[ m1$1*   -1.230  ];
[ m2$1*   -1.240  ];
[ m3$1*    0.344  ];
[ m4$1*   -2.971  ];
[ m5$1*    0.469  ];
[ m6$1*   -0.079  ];
[ m7$1*   -1.043  ];
[ m8$1*    1.456  ];
[ m9$1*   -0.730  ];
[ m10$1*   1.008  ];
[ m11$1*   0.004  ];
[ m12$1*   2.004  ];
[ m13$1*   2.112  ];
[ m14$1*   1.885  ];
[ m15$1*  -1.763  ];
[ m16$1*   1.186  ];

OUTPUT:
TECH1 TECH8;
PLOT: type=plot3;
series = m1 (1) m2 (2) m3 (3) m4 (4) m5 (5) m6 (6) m7 (7) m8 (8)
m9 (9) m10 (10) m11 (11) m12 (12) m13 (13) m14 (14) m15 (15) m16 (16);
! how the variables are presented in the X axis
! (*) separate them by a space
SAVEDATA:
file = data_savedata.txt;
save = cprob;
missflag = 9999;
format = free;
```

A1.0

```
TITLE: Stepwise MIMIC Model DIF Detection
DATA: file is b16-gender.dat;
VARIABLE:
NAMES ARE gender m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16;
MISSING ARE ALL (99);
USEVARIABLES ARE gender m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14
m15 m16;
CATEGORICAL ARE m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16;
CLASSES = c(3);
ANALYSIS:
type=mixture;
```

```
MODEL:
%OVERALL%
[ c#1*-1.291 ];
[ c#2* 0.489 ];

c on gender;
%C#1%
[ m1$1*  -0.612  ];
[ m2$1*  -0.444  ];
[ m3$1*  -1.149  ];
[ m4$1*  -3.141  ];
[ m5$1*  -1.383  ];
[ m6$1*  -1.561  ];
[ m7$1*  -2.091  ];
[ m8$1*  -0.524  ];
[ m9$1*  -0.668  ];
[ m10$1* -1.982  ];
[ m11$1* -0.891  ];
[ m12$1* -0.631  ];
[ m13$1* -0.345  ];
[ m14$1* -0.524  ];
[ m15$1* -1.815  ];
[ m16$1* -1.545  ];

%C#2%
[ m1$1*   3.386   ];
[ m2$1*   2.674   ];
[ m3$1*   0.990   ];
[ m4$1*  -2.473   ];
[ m5$1*   0.490   ];
[ m6$1*   0.404 ];
[ m7$1*  -0.740 ];
[ m8$1*   2.497 ];
[ m9$1*   1.261 ];
[ m10$1*  1.058 ];
[ m11$1*  0.194 ];
[ m12$1*  2.084 ];
[ m13$1*  2.346 ];
[ m14$1*  1.864 ];
[ m15$1* -1.777 ];
[ m16$1*  1.240 ];

%C#3%
[ m1$1*   -1.230 ];
[ m2$1*   -1.240 ];
[ m3$1*    0.344 ];
[ m4$1*   -2.971 ];
[ m5$1*    0.469 ];
[ m6$1*   -0.079];
```

```
[ m7$1*   -1.043];
[ m8$1*    1.456];
[ m9$1*   -0.730];
[ m10$1*   1.008];
[ m11$1*   0.004];
[ m12$1*   2.004];
[ m13$1*   2.112];
[ m14$1*   1.885];
[ m15$1*  -1.763];
[ m16$1*   1.186];
```

A1.1

```
TITLE: Stepwise MIMIC Model DIF Detection
DATA: file is b16-gender.dat;
VARIABLE:
NAMES ARE gender m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16;
MISSING ARE ALL (99);
USEVARIABLES ARE gender m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14
m15 m16;
CATEGORICAL ARE m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16;
CLASSES = c(3);
ANALYSIS:
type=mixture;

MODEL:
%OVERALL%
[ c#1*-1.291 ];
[ c#2* 0.489 ];
c on gender;
m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16 on gender;

%C#1%
[ m1$1*  -0.612  ];
[ m2$1*  -0.444  ];
[ m3$1*  -1.149  ];
[ m4$1*  -3.141  ];
[ m5$1*  -1.383  ];
[ m6$1*  -1.561  ];
[ m7$1*  -2.091  ];
[ m8$1*  -0.524  ];
[ m9$1*  -0.668  ];
[ m10$1* -1.982  ];
[ m11$1* -0.891  ];
[ m12$1* -0.631  ];
[ m13$1* -0.345  ];
[ m14$1* -0.524  ];
[ m15$1* -1.815  ];
[ m16$1* -1.545  ];
m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16 on gender;
```

```
%C#2%
[ m1$1*   3.386   ];
[ m2$1*   2.674   ];
[ m3$1*   0.990   ];
[ m4$1*  -2.473   ];
[ m5$1*   0.490   ];
[ m6$1*   0.404   ];
[ m7$1*  -0.740   ];
[ m8$1*   2.497   ];
[ m9$1*   1.261   ];
[ m10$1*  1.058   ];
[ m11$1*  0.194   ];
[ m12$1*  2.084   ];
[ m13$1*  2.346   ];
[ m14$1*  1.864   ];
[ m15$1* -1.777   ];
[ m16$1*  1.240   ];
m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16 on gender;

%C#3%
[ m1$1*   -1.230  ];
[ m2$1*   -1.240  ];
[ m3$1*    0.344  ];
[ m4$1*   -2.971  ];
[ m5$1*    0.469  ];
[ m6$1*   -0.079  ];
[ m7$1*   -1.043  ];
[ m8$1*    1.456  ];
[ m9$1*   -0.730  ];
[ m10$1*   1.008  ];
[ m11$1*   0.004  ];
[ m12$1*   2.004  ];
[ m13$1*   2.112  ];
[ m14$1*   1.885  ];
[ m15$1*  -1.763  ];
[ m16$1*   1.186  ];
m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16 on gender;
```

A2.0.1

```
TITLE: Stepwise MIMIC Model DIF Detection
DATA: file is data_savedata.txt;
VARIABLE:
NAMES ARE m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16
gender cprob1 cprob2 cprob3 cmod ;
MISSING ARE ALL (9999);
USEVARIABLES ARE m1 cmod gender;
CATEGORICAL ARE m1;
NOMINAL are cmod;
CLASSES = c(3);
ANALYSIS:
```

```
type=mixture;
STARTS=0;
processors = 7;
MODEL:
%OVERALL%
[ c#1*-1.291 ];
[ c#2* 0.489 ];
c on gender;

%C#1%
[cmod#1@2.610 cmod#2@-4.036];
%C#2%
[cmod#1@-3.434 cmod#2@1.739];
%C#3%
[cmod#1@-1.477 cmod#2@-2.631];
```

A2.1.1

```
TITLE: Stepwise MIMIC Model DIF Detection
DATA: file is data_savedata.txt;
VARIABLE:
NAMES ARE m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16
gender cprob1 cprob2 cprob3 cmod ;
MISSING ARE ALL (9999);
USEVARIABLES ARE m3 cmod gender;
CATEGORICAL ARE m3;
NOMINAL are cmod;
CLASSES = c(3);
ANALYSIS:
type=mixture;
STARTS=0;
processors = 7;

MODEL:
%OVERALL%
[ c#1*-1.291 ];
[ c#2* 0.489 ];

c on gender;
m1 on gender;

%C#1%
[cmod#1@2.610 cmod#2@-4.036];
m1 on gender;

%C#2%
[cmod#1@-3.434 cmod#2@1.739];
m1 on gender;

%C#3%
```

```
[cmod#1@-1.477 cmod#2@-2.631];
m1 on gender;

OUTPUT: CINTERVAL;
```

A3.0.

```
TITLE: Stepwise MIMIC Model DIF Detection
DATA: file is b16-gender.dat;
VARIABLE:
NAMES ARE gender m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16;
MISSING ARE ALL (99);
USEVARIABLES ARE gender m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14
m15 m16;
CATEGORICAL ARE m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16;
CLASSES = c(3);
ANALYSIS:
type=mixture;

MODEL:
%OVERALL%
[ c#1*-1.291 ];
[ c#2* 0.489 ];

c on gender;
m9 m12 m13 m15 m16 on gender;

%C#1%
[ m1$1*  -0.612  ];
[ m2$1*  -0.444  ];
[ m3$1*  -1.149  ];
[ m4$1*  -3.141  ];
[ m5$1*  -1.383  ];
[ m6$1*  -1.561  ];
[ m7$1*  -2.091  ];
[ m8$1*  -0.524  ];
[ m9$1*  -0.668  ];
[ m10$1* -1.982  ];
[ m11$1* -0.891  ];
[ m12$1* -0.631  ];
[ m13$1* -0.345  ];
[ m14$1* -0.524  ];
[ m15$1* -1.815  ];
[ m16$1* -1.545  ];
m9 m12 m13 m15 m16 on gender;

%C#2%
[ m1$1*   3.386 ];
[ m2$1*   2.674 ];
[ m3$1*   0.990 ];
[ m4$1*  -2.473 ];
```

```
[ m5$1*   0.490 ];
[ m6$1*   0.404 ];
[ m7$1*  -0.740 ];
[ m8$1*   2.497 ];
[ m9$1*   1.261];
[ m10$1*  1.058];
[ m11$1*  0.194 ];
[ m12$1*  2.084   ];
[ m13$1*  2.346   ];
[ m14$1*  1.864   ];
[ m15$1* -1.777   ];
[ m16$1*  1.240   ];
m9 m12 m13 m15 m16 on gender;

%C#3%
[ m1$1*   -1.230  ];
[ m2$1*   -1.240  ];
[ m3$1*    0.344  ];
[ m4$1*   -2.971  ];
[ m5$1*    0.469  ];
[ m6$1*   -0.079  ];
[ m7$1*   -1.043  ];
[ m8$1*    1.456  ];
[ m9$1*   -0.730  ];
[ m10$1*   1.008  ];
[ m11$1*   0.004  ];
[ m12$1*   2.004  ];
[ m13$1*   2.112  ];
[ m14$1*   1.885  ];
[ m15$1*  -1.763  ];
[ m16$1*   1.186  ];
m9 m12 m13 m15 m16 on gender;
```

A.4.1.

```
TITLE: Stepwise MIMIC Model DIF Detection
DATA: file is b16-gender.dat;
VARIABLE:
NAMES ARE gender m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16;
MISSING ARE ALL (99);
USEVARIABLES ARE gender m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14
m15 m16;
CATEGORICAL ARE m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16;
CLASSES = c(3);
ANALYSIS:
type=mixture;

MODEL:
%OVERALL%
[ c#1*-1.291 ];
[ c#2* 0.489 ];
```

```
c on gender;
m12 m13 m15 m16 on gender;

%C#1%
[ m1$1*  -0.612  ];
[ m2$1*  -0.444  ];
[ m3$1*  -1.149  ];
[ m4$1*  -3.141  ];
[ m5$1*  -1.383  ];
[ m6$1*  -1.561  ];
[ m7$1*  -2.091  ];
[ m8$1*  -0.524  ];
[ m9$1*  -0.668  ];
[ m10$1* -1.982  ];
[ m11$1* -0.891  ];
[ m12$1* -0.631  ];
[ m13$1* -0.345  ];
[ m14$1* -0.524  ];
[ m15$1* -1.815  ];
[ m16$1* -1.545  ];
m12 m13 m15 m16 on gender;

%C#2%
[ m1$1*   3.386   ];
[ m2$1*   2.674   ];
[ m3$1*   0.990   ];
[ m4$1*  -2.473   ];
[ m5$1*   0.490   ];
[ m6$1*   0.404   ];
[ m7$1*  -0.740   ];
[ m8$1*   2.497   ];
[ m9$1*   1.261   ];
[ m10$1*  1.058   ];
[ m11$1*  0.194   ];
[ m12$1*  2.084   ];
[ m13$1*  2.346   ];
[ m14$1*  1.864   ];
[ m15$1* -1.777   ];
[ m16$1*  1.240   ];
m12 m13 m15 m16 on gender;

%C#3%
[ m1$1*   -1.230  ];
[ m2$1*   -1.240  ];
[ m3$1*    0.344  ];
[ m4$1*   -2.971  ];
[ m5$1*    0.469  ];
[ m6$1*   -0.079  ];
[ m7$1*   -1.043  ];
```

```
[ m8$1*    1.456  ];
[ m9$1*   -0.730  ];
[ m10$1*   1.008  ];
[ m11$1*   0.004  ];
[ m12$1*   2.004  ];
[ m13$1*   2.112  ];
[ m14$1*   1.885  ];
[ m15$1*  -1.763  ];
[ m16$1*   1.186  ];
m12 m13 m15 m16 on gender;
```