International Journal of Assessment Tools in Education
2024, Vol. 11, No. 2, 249–269

https://doi.org/10.21449/ijate.1387041

journal homepage: https://dergipark.org.tr/en/pub/ijate https://ijate.net/ ***Research Article***

Detection of differential item functioning with latent class analysis: PISA 2018 mathematical literacy test

Selim Daşçıoğlu[1*], Tuncay Öğretmen[1]

[1]Ege University, Faculty of Education, Department of Educational Sciences, İzmir, Türkiye

ARTICLE HISTORY

Received: Nov. 06, 2023

Accepted: Feb. 20, 2024

Keywords:

Cross country DIF,
Finite mixture models,
MIMIC,
Latent class analysis,
PISA.

Abstract: The purpose of this research is to determine whether PISA 2018 mathematical literacy test items show a differential item functioning across countries. For this purpose, only the items in booklet number three were examined using the MIMIC method with Latent Class Analysis (LCA) approach. PISA 2018 tests are mostly developed in English. Therefore, in DIF analyses, the reference group is the UK, while the focal groups consist of the other countries examined in the research (Türkiye, Finland, Japan, and the USA). According to the results, of the 23 test items, statistically significant DIF was observed in eight items in the UK-Türkiye sample, in seven items in the UK-Finland sample, in eleven items in the UK-Japan sample, and in three items in the UK-USA sample. It is seen that the effect and size of DIF in non-homogeneous groups differ between groups and these effects can be examined in more detail with the LCA method.

1. INTRODUCTION

The emerging technological developments and globalization offer countries the opportunity to develop their educational policies in a way that can help them keep up with the changing world and direct those changes. Large-scale international exams and practices also provide an opportunity for countries to measure their own levels and compare the results with those of other countries. One of these applications, the Program for International Student Assessment (PISA), is a program implemented by the Organization for Economic Co-operation and Development (OECD) and aims to measure the ability of 15-year-old students to utilize their reading comprehension, mathematics, scientific knowledge, and skills to cope with real-life problems. International monitoring research in education enables countries to assess their situation, compare their level with that of other countries, and make social and political decisions accordingly (MEB, 2019).

Considering that such decisions would be taken based on the measurement results, the quality of the measurement tools becomes important. One of the most important features of a measurement tool is its validity. In its broadest sense, validity is the degree to which measurement results serve the purpose (Nunnally & Bernstein, 1994). For this, all test items are

*CONTACT: Selim Daşçıoğlu ✉ selimdascioglu@gmail.com Ege University, Faculty of Education, Department of Educational Sciences, İzmir, Türkiye

e-ISSN: 2148-7456

expected to distinguish individuals well. Zumbo (1999) stated that the concept, method, and process of validation are at the core of measurement, and in the absence of validity studies, the inferences to be made from the measurement results will be meaningless.

Validity is not related to measurement results but to inferences made from measurement results (Zumbo, 1999). From this point of view, based on the results of international tests, it is necessary to emphasize the validity of making valid comparisons and inferences between countries.

Sometimes, the results obtained from the tests may vary according to the subgroups of the individuals. Differential item functioning (DIF) occurs when test takers from different subgroups show different success probabilities on the item after matching the basic ability that the item aims to measure (Camilli & Shepard 1994; Clauser & Mazor, 1998; Zumbo, 1999). Contrarily, item bias occurs when the probability of answering an item correctly differs for individuals at the same ability level but from different subgroups. This is due to a factor other than the characteristic the test item is intended to measure (Camilli & Shepard 1994; Clauser & Mazor, 1998; Zumbo, 1999). Accordingly, biased items show DIF. However, not every item showing DIF may be biased. Therefore, bias is a systematic error that affects the inferences made from the measurement results (Zumbo, 1999). In comparisons between subgroups such as gender or countries according to test results, it is important for test developers and policymakers to determine whether test items show DIF in terms of the relevant variable to make more valid comparisons and more unbiased measurements.

Additionally, there are more complex structural equation models that include many latent or observed variables and covariates and aim to determine the relationships between these variables. In such models, if DIF or direct effects arising from the covariate are predetermined and not included in the established model, biased results may occur (Vermunt, 2010).

Individuals can be divided into observable subgroups like gender, religion, language, race, and socioeconomic level. Additionally, individuals can be divided into subgroups that cannot be directly observed according to some latent traits like intelligence, achievement, attitude, alcohol addiction, etc., that we are trying to measure. LCA is a statistical method that allows the categorization of individuals into meaningful latent classes for the measured latent trait (Lanza & Collins, 2010; McCutcheon, 1987). DIF can occur between observed groups and latent classes. Especially in cases where the observed groups are not homogeneous, ignoring latent classes may lead to biased results and biased decisions (Sawatzky et al., 2018). Therefore, finite mixture models have been developed that allow the DIF to be among the latent classes and the observed groups.

Most of the tests in PISA 2018 were developed in English and French, and cross-cultural and cross-linguistic adaptations were made by relevant stakeholders (OECD, 2016c). However, no matter how meticulously the cross-cultural adaptation is carried out, mistakes can be made that will cause psychometric bias. Considering that these tests aim to measure latent structures, it is better to realize how difficult this task is. It is a process that does not expire and must be repeated at regular intervals (Messick, 1989). Therefore, conducting all validity studies, such as bias studies, during test development and adaptation processes and at the end of the actual application will contribute to future applications and processes.

One of the constructs that PISA aims to measure is mathematical literacy. OECD (2019) defined mathematical literacy as:

> Mathematical literacy is an individual's capacity to reason mathematically and to formulate, employ, and interpret mathematics to solve problems in a variety of real-world contexts. It includes concepts, procedures, facts and tools to describe, explain and predict phenomena. It assists individuals to know the role that mathematics plays in the world and to make the well-founded judgments and decisions needed by constructive, engaged and reflective 21st century citizens.

As it can be understood from the definition, the measurement of high-level mathematical skills related to real life and the development of skills based on the measurement results are of great importance in terms of raising individuals with these skills.

Considering all this information, it becomes necessary to conduct a DIF research to make meaningful comparisons between countries based on the results of PISA 2018 literacy tests. In addition to the country variable, which is the observed group variable while conducting this research, LCA will be used in order to consider the subgroups of individuals according to the latent feature of mathematical literacy.

In this study, the primary reason for using the MIMIC model with the LCA approach, as suggested by Masyn (2017), is to test whether the items show DIF according to the covariate, stepwise. In other words, the three-step procedure is used. With the addition of the covariate to the latent class model, the item response probability of individuals changes, and therefore, the latent class membership of some individuals may also change (Vermunt, 2010). However, this is undesirable in the current research. Because it is thought that the covariate (country variable in this study) is not a predictor of the latent class variable (mathematical literacy in this study). The three-step procedure will enable controlling this undesirable situation in the second step (analysis steps will be explained later), in which it is determined whether the items show DIF (Vermunt, 2010; Masyn, 2017). Furthermore, there is a limited number of studies to determine DIF with the latent class MIMIC method in the literature (Masyn, 2017; Tsaousis et al., 2020). It is thought that this study, which is based on real data, will contribute to the literature.

In this study, whether the mathematical literacy subtest items in PISA 2018, in which Türkiye and many OECD countries participated, show DIF across countries will be examined with the latent class MIMIC method. In this cross-country research, the other countries within the scope of the research will be compared in pairs with the UK, which is the reference group, given that the OECD is Europe-based and the languages in which the test was developed are English and French. While choosing other countries, attention was paid to the fact that these countries were from different parts of the world and from different cultures, and therefore Türkiye, Finland, Japan and the USA were determined.

Adapting a test that aims to measure a latent construct for other cultures is a very complicated and difficult process. This process aims to keep the validity and reliability of the measurements high with many quantitative and qualitative research techniques (Hambleton, Merenda & Spielberger, 2005). This is also true for international applications such as PISA, the Progress in International Reading Literacy Study (PIRLS), and the Trends in International Mathematics and Science Study (TIMSS). It is thought that DIF determination and bias studies will shed light on test development and adaptation studies and will contribute to increasing the validity of the decisions to be taken according to the test results.

2. METHOD

The target population of PISA is students between 15 years and three months and 16 years and two months who are in seventh grade and above, attending educational institutions located in the participating countries (OECD, 2016a). Approximately 600,000 students from 79 countries, 37 of which are OECD members, participated in PISA 2018 application. This sample represents the target population of approximately 32 million students (MEB, 2019).

Only the UK, Türkiye, Japan, Finland and the USA samples were analyzed in this study. Additionally, it was limited to booklet number three of the mathematical literacy test. Here is a country-wise bifurcation of 1442 participating students: UK 516, Türkiye 281, Japan 243, Finland 217, and USA 185.

Research data were obtained from the official website of OECD (https://www.oecd.org/pisa/data/), which prepared the PISA 2018 application. The application, scoring, and coding of the mathematical literacy test examined in the research were carried out by the relevant stakeholders (OECD, 2016b).

Only the mathematical literacy subtest of booklet number three of the PISA 2018 was examined in a recent study. There are 23 items in the test. Item 22, with a partially correct answer (0-1-2), was divided into two categories (0-1), with fully correct answers as "1" and other answers as "0". All other items are in two categories. Among the students who took the test, those who could not answer at least one of the test items (156 response patterns) because they could not see the test period or for any other reason were excluded from the data set. Apart from these, students who saw the question and left it blank were coded as "0", assuming that they did not answer because they did not know the correct answer.

Of the 1286 students remaining at the end of these procedures, 451 were from the UK, 253 from Türkiye, 188 from Finland, 226 from Japan, and 168 from the USA. The descriptive statistics of the test are given in the table below. Analyses were made with the TAP (Test Analysis Program).

Table 1. *Descriptive statistics of the PISA 2018 mathematical literacy test.*

Test Statistics		Test Statistics	
Number of Students	1286	Variance	22.26
Number of Items	23	Skewness	0.10
Lowest Score	0	Kurtosis	-0.67
Highest Score	23	Mean of Item Difficulty	0.47
Median	11	Mean of Item Discrimination	0.49
Mean	10.74	Mean of Item Point Biserial Discrimination	0.39
Standard deviation	4.72	KR-20	0.84

When the values in Table 1 are examined, it is seen that the group is heterogeneous, the distribution is slightly flat (-0.67), and the skewness (0.10) is close to zero. Based on this information, it can be assumed that the distribution is normal (Fraenkel et al., 2011). In addition, it can be said that the test has medium difficulty according to the mean item difficulty index (0.47), and the test distinguishes the upper group and the lower group from each other well according to the mean point double series discrimination values (0.39). Furthermore, according to the alpha coefficient (0.84), the reliability of the test in terms of internal consistency is high (Kerlinger, 1999).

Confirmatory factor analysis was performed with the R package lavaan for the model in which all items of the mathematical literacy test were collected in a single factor (mathematical literacy) structure and used diagonally weighted least squares (DWLS) estimation. Because chi-square (χ^2 = 385.615, *sd* = 230, *p*<0.001) was affected by the sample size and tended to be statistically significant, other goodness-of-fit values were examined. According to the analysis results, the RMSEA (0.023), CFI (0.990), TLI (0.989), GFI (0.982), and AGFI (0.979) values indicated a good model fit; the SRMR (0.060) value gave an acceptable model fit value. Therefore, it can be accepted that the test measures a single-factor construct (Harrington, 2009).

2.1. Latent Class Analysis

In 1950, Lazarsfeld performed a cluster analysis with data consisting of dichotomous items. In 1974, Goodman developed this analysis using the method of maximum likelihood estimation with categorical variables and made it applicable in practice (Magidson & Vermunt, 2004).

LCA is a statistical method for detecting and describing homogeneous and not directly observable (latent) subgroups in which individuals are separated according to a certain latent characteristic. This method comprises only one subgroup in which each individual is included. These subgroups of individuals cannot be known precisely due to measurement error. Additionally, the responses of individuals in each latent class to the indicator variables are independent of each other. This is called the local independence assumption, which is the only assumption of this model. LCA is used in a wide range of fields, such as behavioral sciences, medicine, education and social sciences, and economics (Magidson & Vermunt, 2004).

Iterative methods such as expectation-maximization or the Newton-Raphson algorithm are used in parameter estimation in LCA (Lanza & Collins, 2010; Magidson & Vermunt, 2004; McCutcheon, 1987).

We can divide the selection of the most suitable model in LCA into two: absolute model fit and comparative model fit (Lanza & Collins, 2009). If there are a certain number of latent classes expected for the latent class variable according to the theoretical background, absolute model fit can be used. In absolute model fit, the likelihood ratio chi-square goodness-of-fit statistic, the G2 test (shown as L2 in Latent Gold software) is used (Magidson & Vermunt, 2004).

H_0 tested here is, "There is no statistically significant difference between the selected model and the population distribution." In order for H_0 to be accepted and selected as the appropriate model, $p>\alpha$ is expected in the determined K-class model. Otherwise, if ($p<\alpha$), H_0 is rejected, and the K-class model determined according to this statistic cannot be used, or other model fit methods can be used (Lanza &Collins, 2010; McCutcheon, 1987). However, as the number of indicator variables in the model and the number of categories of these variables increase, and as gaps occur in the cells in the contingency table, sparseness will occur, and G^2 will tend to be higher (Lanza & Collins, 2010; McCutcheon, 1987). In this case, this method, which is desired to be used for model selection, can be misleading. In such a case, the use of comparative model selection may be healthier.

One of the statistics used in the comparative model selection is the G^2 difference (ΔG^2) statistic. In this method, the G^2 differences of two models with class K and class (K+1) are tested. However, we cannot directly test two models with different latent class numbers in this way because we cannot know the correct reference distribution. Therefore, the bootstrap method is used for both models, and then ΔG^2 is tested. Here, H_0 means that there is no significant difference between the K-class model and the (K+1) class model, therefore, if $p>\alpha$, the K-class model, with a lower number of parameters and a simpler one, is chosen based on the principle of parsimony. Otherwise, if ($p<\alpha$), it is seen that there is a significant difference in the (K+1) class model; the (K+1) class model is selected (Magidson & Vermunt, 2004).

Other statistics most frequently used in comparative model fit are information criteria. These are information criteria such as BIC (Bayesian Information Criterion), AIC (Akaike Information Criterion), and CAIC (Consistent Akaike Information Criterion). When comparing models with these information criteria, the model with the lower information criterion value is preferred to the model with the higher value (Lanza & Collins, 2009; McCutcheon, 1987; Magidson & Vermunt, 2004).

It should also be added that whether the appropriate model is chosen by one of the absolute or comparative model fit methods when the latent classes are examined (the responses of the individuals in the latent classes to the indicator variables and the predicted item-response probabilities), the classes should be well separated from each other and well defined (Lanza & Collins, 2010; McCutcheon, 1987; Vermunt & Magidson, 2002). If the latent classes in the selected model are not homogeneous enough and cannot be separated from each other in a meaningful way, i.e., they cannot be defined well, it will not make sense for the applied statistics to point to the selected model.

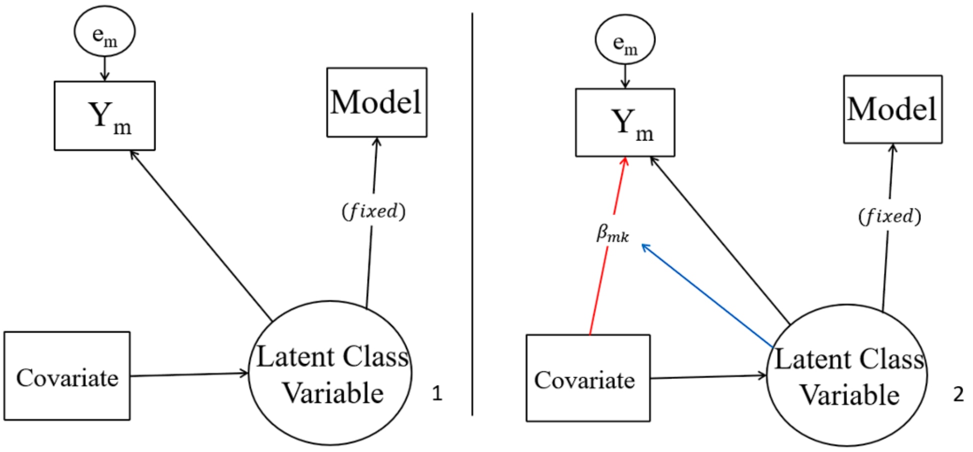
2.2. Latent Class MIMIC Model Steps

In this study, the steps of the latent class MIMIC method proposed by Masyn (2017) will be used. The steps of the analysis are as follows:

Initial Stage (Step 0): At this stage, LCA is performed with indicator variables (test items in this study) without a covariate, and the most suitable K-class model is determined. Individuals assigned to classes according to the selected K-class model are then numbered according to these classes. This is the first step of the three-step approach proposed by Vermunt (2010). The reason for class enumeration before including the covariate in the model is that when the covariate is included, the changes that may occur in the item response probabilities and latent class memberships of some individuals cannot be ignored (Nylund-Gibson et al., 2014).

Step 1: In this step, two different models are estimated. In the first model (M1.0), the group variable is included as a covariate in addition to the initial model (K-class). Here, the group variable has a direct effect only on the latent class variable. In other words, this model can be called the No-DIF model. In the second model (M1.1), the covariate included in the model has a direct effect on both items (non-uniform DIF where the effects of the covariate on the items are released to vary between classes). So, this model can also be called All-DIF. Then, the two models are compared using the likelihood ratio test (LRT). In this comparison, twice the difference of the loglikelihood (Δ-2LL) values of the models is tested with the chi-square test, which considers the difference in the number of parameters (ΔNpar) of the models as degrees of freedom. If H_0 cannot be statistically rejected (p>α), there is no evidence that the covariate is a source of DIF, and the analysis ends there. However, if H_0 is rejected (p≤α), there is sufficient evidence that the covariate can be a source of DIF for at least one of the indicator variables, and the second step is taken.

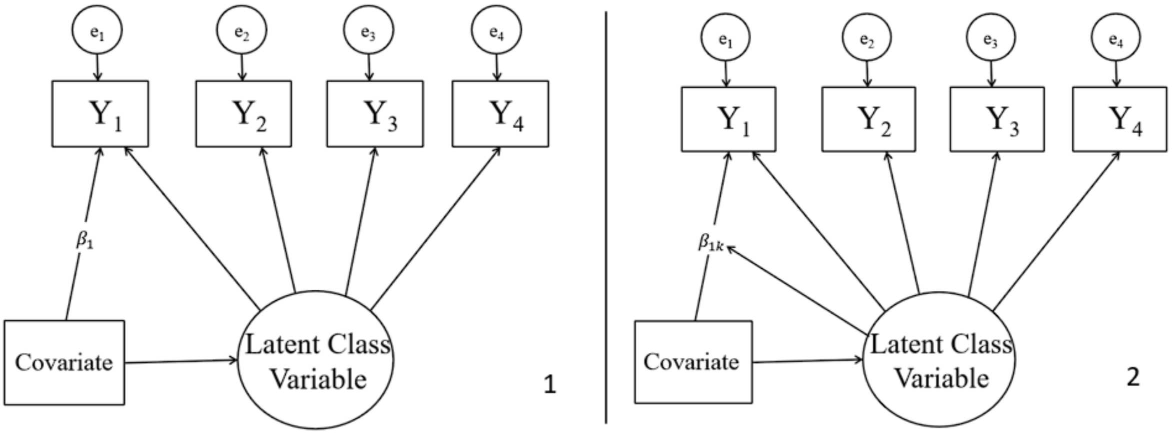
Figure 1. *In step 2, with a three-step approach, the M2.0.m model (1) in which the item Y_m covariate has no DIF effect, and the M2.1.m model (2) in which the covariate has a non-uniform DIF effect (β_{mk} is log odds ratio of endorsing item Y_m given membership latent class k for one-unit positive difference of covariate).*



Step 2: In this step, a non-uniform DIF test will be performed for each indicator variable (test item) separately. The three-step method is used for this (Magidson & Vermunt, 2004). In the first model, their membership in the initially obtained K-class model is fixed. The first item (m_1) and the covariate are included in the model. Here, the covariate has no direct effect on the item (Figure 1 left (1)), and this model is shown as M2.0.1. In the second model (M2.1.1), the covariate has a direct effect on the item, and the effects of the covariate on the items were left free to change between classes (Figure 1 right (2)). Then, the two models are compared with LRT. These model comparisons are made separately for each item (For example, M2.0.2 and M2.1.2 models for item 2). For items with statistically significant pairwise comparisons, there is sufficient evidence for DIF resulting from the covariate.

Step 3: In this step, a new latent class MIMIC model is estimated in line with the findings from the second step. In the model (M3.0), there is a non-uniform DIF effect for the items whose DIF was determined in step 2. For items for which no evidence of DIF can be obtained, the covariate has no direct effect. M3.0 is compared in pairs with M1.0 and M1.1. As a result of this comparison, it is expected that M3.0 is statistically better (p<α) than M1.0 and not worse than M1.1 (p>α).

Figure 2. *Models where (1) in the latent class model, the covariate is the source of uniform DIF for the indicator variable Y_1 and (2) the source of non-uniform DIF for the same variable.*

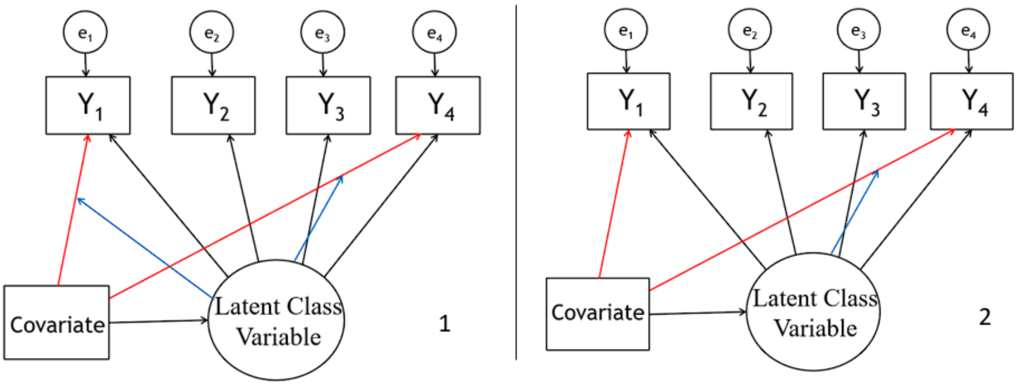


Step 4: In this step, a uniform DIF test will be performed for items showing DIF. In the estimated model (M4.1), unlike the M3.0 model, in one of the items showing non-uniform DIF, the variation of the common effect between the latent classes is fixed (Figure 2). Thus, only the covariate and the latent variable have a direct effect on that item. These models are set up separately for each item. In each model, the uniform DIF effect of only one item is tested (Figure 3). Each model is then individually compared (LRT) to the M3.0. If new models (such as M4.1 and M4.2) are not statistically worse than M3.0 ($p>\alpha$), there is uniform evidence of DIF. Conversely, it can be said that there is evidence of non-uniform DIF.

Step 5: If there are items with uniform DIF detected in Step 4, a new model is estimated (M5.0) that these items show uniform DIF, unlike M3.0. Then, M3.0 LRT is compared with this model, and it is expected that M5.0 is not statistically worse ($p> \alpha$).

Step 6: The direction and effect size of DIF in items with DIF will be determined. For this, the estimated coefficient (β) of the direct effect from the country variable to the item in items showing uniform DIF will be examined. As the value for the UK is coded as 0 and that for the other country is coded as 1, a positive coefficient will indicate DIF in favor of the other country (focal group), and a negative will indicate DIF in favor of the UK (reference group). For items with non-uniform DIF, the latent class or classes and the direction of the DIF will be measured. When evaluating the effect size, according to the ETS (Educational Testing Service) criteria, those equal to or less than 0.44 will be considered negligible (small) DIFs, those greater than 0.64 will be considered large DIFs, and those between these values will be considered medium-sized DIFs (Masyn, 2017; Tsaousis et al., 2020). Analyses were made with the Latent Gold 5.1 program (Vermunt & Magidson, 2016).

Figure 3. *In step 4 for Y_1; (1) non-uniform DIF effect for Y_1 and Y_4 (like M3.0), (2) uniform DIF for Y_1 and non-uniform DIF for Y_4 (like M4.1).*



3. FINDINGS

In the initial stage (Step 0), LCA was performed with mathematical literacy items in each sample. The following table shows the results of the LCA.

Table 2. *LCA results by samples.*

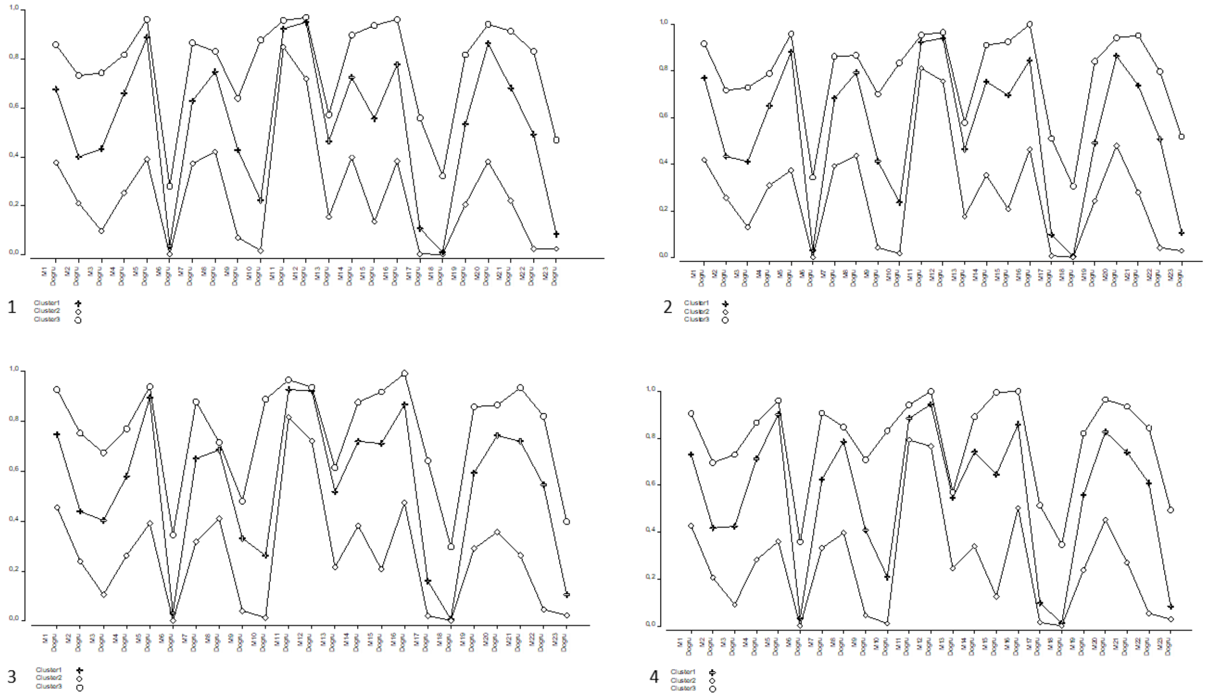
Sample	Model	LL	BIC(LL)	AIC(LL)	AIC3(LL)	CAIC(LL)	Npar	L^2	*df*	*p*
UK - Türkiye	1 Class	-9193.03	18536.86	18432.06	18455.06	18559.86	23	9207.16	681	0.00
	2 Class	-8164.76	16637.68	16423.51	16470.51	16684.68	47	7150.61	657	0.00
	3 Class	-7988.66	16442.85	16119.32	16190.32	16513.85	71	6798.43	633	0.00
	4 Class	-7920.36	16463.61	16030,72	16125.72	16558.61	95	6661.82	609	0.00
UK - Finland	1 Class	-8243.20	16634.97	16532.40	16555.40	16657.97	23	8282.64	616	0.00
	2 Class	-7415,20	15134.01	14924.40	14971.40	15181.01	47	6626.64	592	0.00
	3 Class	-7269.18	14997.00	14680.35	14751.35	15068,00	71	6334.59	568	0.00
	4 Class	-7199.01	15011.71	14588.02	14683.02	15106.71	95	6194.26	544	0.00
UK - Japan	1 Class	-8930.37	18010.64	17906.73	17929.73	18033.64	23	9070.12	654	0.00
	2 Class	-8040.66	16387.65	16175.32	16222.32	16434.65	47	7290.71	630	0.00
	3 Class	-7890.95	16244.66	15923,90	15994,90	16315.66	71	6991.29	606	0.00
	4 Class	-7833,30	16285.77	15856.59	15951.59	16380.77	95	6875.98	582	0.00
UK - USA	1 Class	-7968.92	16085.69	15983.84	16006.84	16108.69	23	8015.21	596	0.00
	2 Class	-7139.89	14581.91	14373.79	14420.79	14628.91	47	6357.16	572	0.00
	3 Class	-7005.27	14466.93	14152.53	14223.53	14537.93	71	6087,90	548	0.00
	4 Class	-6940.65	14491.98	14071.31	14166.31	14586.98	95	5958.68	524	0.00

Npar: number of parameters, df: degrees of freedom

Models with more than four latent classes are not given in Table 2 because they are not well defined. Table 2 shows that the p values of all latent class models are statistically significant. However, as stated under the heading "Parameter Estimation and Model Selection", because this figure tends to be statistically significant due to the number of variables and sparseness, other information criteria will be used in the model selection. Considering the models with the lowest information criterion values in all samples, the three-class model, according to BIC and CAIC; according to AIC and AIC3, the four-class model is more suitable for the data. Güngör Culha (2012) concluded in his research that "BIC and CAIC criteria give better results than other criteria in making the right decision while choosing the most suitable model as the sample grows." Additionally, when the latent classes in three-class and four-class models are examined, it is seen that the classes are more homogeneous in the former model. Based on this information, it was concluded that the most suitable model for the data in all samples was the three-class model. Figure 4 shows the item-response probabilities of the latent classes in three-class models.

Figure 4 shows that the latent classes are separated from each other for all samples. The class with the highest probability of rendering a correct answer for all items was named as "High Achiever Class (HAC)", the lowest class as "Low Achiever Class (LAC)" and the other class as "Moderate Achiever Class (MAC)". The sizes of the latent classes are as follows: 15.4% of the UK–Türkiye sample is in HAC, 43.5% in MAC, and 41.1% in LAC. Of the UK–Finland sample, 14.7% are in HAC, 47.9% in MAC and 37.4% in LAC. Of the UK-Finland sample, 14.7% are in HAC, 47.9% in MAC, and 37.4% in LAC. Of the UK-Japan sample, 18.1% were in HAC, 46.1% in MAC, and 34.8% in LAC. In the UK-USA sample, 11.2% are in HAC, 49.6% in MAC and 39.2% in LAC. From here, we move on the next step of the analysis.

Figure 4. *Item-response probabilities of latent classes in the three-class model: (1) UK-Türkiye, (2) UK-Finland, (3) UK-Japan, (4) UK-USA.*



MIMIC analysis results of the UK-Türkiye, UK-Finland, UK-Japan, and UK-USA samples are given in Appendices (Table 5, 6, 7 and 8), respectively. In the first step, the M1.0 No-DIF model was compared with the M1.1 All-DIF model using LRT. When the tables were examined, a statistically significant difference was observed between the models in all samples (*Δ-2LL*=263.21, *ΔNpar*=69, *p*<0.001 for the UK-Türkiye; *Δ-2LL*=246.06, *ΔNpar* = 69, *p*<0.001 for the UK-Finland; *Δ-2LL*=428.34, *ΔNpar*=69, *p*<0.001 for the UK-Japan, and *Δ-2LL*=108.35, *ΔNpar*=69, *p*<0.001 for the UK-USA). This is sufficient proof that the country variable is a source of DIF for at least one of the indicator variables in at least one of the latent classes. From this point of view, the second step was started.

In step 2, the no DIF model (M2.0.m) established for each item and the non-uniform DIF model (M2.1.m) were compared with the LRT. In comparisons with statistically significant difference between them, it was concluded that the relevant item contained DIF originating from the country variable. According to the results in the tables in Appendices: In the UK-Türkiye sample, in items 5, 7, 11, 13, 14, 15, 16, 20, 21 and 22; in the UK-Finland sample, items 1, 7, 8, 11, 13, 15, 19 and 22; DIF originating from the country variable was found in items 1, 4, 8, 9, 11, 12, 13, 15, 17, 20 and 23 in the UK-Japan sample and in items 10, 16 and 23 in the UK-USA sample.

In step 3, a new model was estimated (M3.0), in which there was a non-uniform effect of DIF on the items in which DIF was detected in the previous step, and there was no direct effect on the other items from the country variable (M3.0), and this model was compared with M1.0 and M1.1. As expected in the UK-Finland and the UK-USA samples, M3.0 was statistically better (*p*<0.05) than M1.0 and not statistically worse than M1.1 (*p*>0.05). However, there was a statistically significant difference between the M3.0 and M1.0 and M1.1 models in the UK-Türkiye and the UK-Japan samples. Then, the BIC values of the models were examined. In both samples, the BIC of M3.0 was considerably lower than the BIC of M1.1 (in the UK-Türkiye, BIC= 16442.78 in M3.0, BIC=16627.33 in M1.1; in the UK-Japan, BIC= 16083.68 at M3.0, BIC=16259.45 at M1.1). Based on this information, it was decided that the most appropriate latent class MIMIC model up to this stage was M3.0 in all samples (Masyn, 2017; Tsaousis et al., 2020).

In step 4, the DIF type of the items for which DIF was detected in previous steps will be determined. For this, the variation of the direct effect between latent classes in one of these items at a time was consistent across classes (uniform DIF model for the item). The estimated models were compared with the M3.0. A statistically significant difference was accepted as evidence that the relevant item contained non-uniform DIF, and otherwise, it contained uniform DIF. Accordingly, in the UK-Türkiye sample, uniform DIF caused by the country variable was found in items 5, 7, 11, 14, 21, and 22, and non-uniform DIF in items 13, 15, 16, and 20. In the UK-Finland sample, items 1, 7, 8, 19, and 22 are uniform caused by the country variable, non-uniform in items 11, 13, and 15; In UK-Japan sample, items 1, 4, 9, 11, 13, 17, 20, and 23 are uniform caused by the country variable, and non-uniform for items 8, 12, and 15; In the UK-USA sample, uniform DIF was detected in items 10 and 16 caused by the country variable, and non-uniform DIF in item 23.

In step 5, a new latent class MIMIC model (M5.0) was estimated with the information obtained in the previous step, in which the items showing DIF had a direct effect from the country variable according to the type of DIF and the other items had no direct effect from the country variable. M5.0 and M3.0 were compared with LRT, and there was no statistically significant difference between models in all samples. In other words, M5.0 can be considered the most suitable model for all samples.

In step 6, the direction and magnitude of the DIF effects arising from the country variable in the items were examined. The results are shown in Table 3. According to the results in Table 3, a statistically significant, uniform, and negligible DIF effect caused by the country variable was observed in items 7, 11, 14, 21, and 22 in the UK-Türkiye sample. The DIF effect in items 7, 11, and 14 is in favor of Türkiye, but the DIF effect in items 21 and 22 is in favor of the UK. When the items showing non-uniform DIF caused by the country variable were examined, it was observed that for item 13, DIF was small in LAC and medium in MAC, a statistically significant, and DIF in favor of the UK. In items 15 and 16, the DIF effect, which is statistically significant only in MAC, is in favor of the UK and of negligible magnitude. In item 20, the DIF effect, which is statistically significant only in LAC, is in favor of the UK and is of negligible magnitude.

In the UK-Finland sample, items 1, 7, 8, 19, and 22 showed a statistically significant, uniform, and negligible DIF effect caused by the country variable. While the DIF effect in items 1, 7, and 8 is in favor of Türkiye, the DIF effect in items 19 and 22 is in favor of the UK. In item 13, the statistically significant non-uniform DIF effect caused by the country variable in favor of the UK is moderate in LAC and negligible in MAC. The statistically significant DIF effect in item 15 is in favor of Finland and negligible in LAC and MAC.

In the UK-Japan sample, items 1, 4, 9, 11, 13, 17, 20, and 23 showed statistically significant uniform DIF caused by the country variable. While the effect size of DIF in items 9 and 20 was medium, it was observed that DIF was negligible in other items. In addition, while the DIF effect in items 1, 11, and 17 is in favor of Japan, it is in favor of the UK in items 4, 9, 13, 20, and 23. The non-uniform DIF effect, which is statistically significant in item 8, is in favor of the UK and negligible in MAC and HAC. In item 12, the DIF effect, which is statistically significant only in MAC, is in favor of the UK and is negligible. In item 15, the statistically significant DIF effect is in favor of Japan and negligible in LAC and MAC. Another issue seen in Table 3 is that although item 5 in the UK-Türkiye sample and item 11 in the UK-Finland sample showed DIF in the previous steps, the DIF effects are not statistically significant in M5.0.

Table 3. *DIF effects from country variable in M5.0.*

		C1 (LAC)			C2 (MAC)			C3 (HAC)		
Samples	Item	β	SE	*p*	β	SE	*p*	β	SE	*p*
UK - Türkiye	5	0.11	0.06	0.06	0.11	0.06	0.06	0.11	0.06	0.06
	7	0.09	0.04	0.04	0.09	0.04	0.04	0.09	0.04	0.04
	11	0.38	0.09	0.00	0.38	0.09	0.00	0.38	0.09	0.00
	13	-0.28	0.11	0.01	-0.46	0.09	0.00	-0.04	0.10	0.70
	14	0.11	0.05	0.02	0.11	0.05	0.02	0.11	0.05	0.02
	15	-0.02	0.10	0.85	-0.32	0.07	0.00	-0.31	0.25	0.22
	16	-0.10	0.07	0.15	-0.33	0.08	0.00	-1.16	1.03	0.26
	20	-0.25	0.08	0.00	0.07	0.11	0.53	0.03	0.26	0.92
	21	-0.25	0.05	0.00	-0.25	0.05	0.00	-0.25	0.05	0.00
	22	-0.28	0.06	0.00	-0.28	0.06	0.00	-0.28	0.06	0.00
UK - Finland	1	0.19	0.06	0.00	0.19	0.06	0.00	0.19	0.06	0.00
	7	0.26	0.05	0.00	0.26	0.05	0.00	0.26	0.05	0.00
	8	0.17	0.05	0.00	0.17	0.05	0.00	0.17	0.05	0.00
	11	0.16	0.10	0.13	1.49	1.83	0.42	1.08	1.84	0.56
	13	-0.53	0.20	0.01	-0.41	0.08	0.00	0.07	0.14	0.60
	15	0.31	0.09	0.00	0.20	0.09	0.02	-0.94	0.93	0.31
	19	-0.16	0.05	0.00	-0.16	0.05	0.00	-0.16	0.05	0.00
	22	-0.22	0.06	0.00	-0.22	0.06	0.00	-0.22	0.06	0.00
UK - Japan	1	0.18	0.05	0.00	0.18	0.05	0.00	0.18	0.05	0.00
	4	-0.28	0.05	0.00	-0.28	0.05	0.00	-0.28	0.05	0.00
	8	0.10	0.08	0.19	-0.33	0.07	0.00	-0.33	0.13	0.01
	9	-0.52	0.07	0.00	-0.52	0.07	0.00	-0.52	0.07	0.00
	11	0.31	0.09	0.00	0.31	0.09	0.00	0.31	0.09	0.00
	12	-0.11	0.08	0.17	-0.37	0.11	0.00	-1.24	1.60	0.44
	13	-0.10	0.05	0.02	-0.10	0.05	0.02	-0.10	0.05	0.02
	15	0.23	0.10	0.02	0.30	0.10	0.00	-1.09	0.94	0.25
	17	0.34	0.06	0.00	0.34	0.06	0.00	0.34	0.06	0.00
	20	-0.49	0.06	0.00	-0.49	0.06	0.00	-0.49	0.06	0.00
	23	-0.15	0.07	0.04	-0.15	0.07	0.04	-0.15	0.07	0.04
UK - USA	10	-0.32	0.09	0.00	-0.32	0.09	0.00	-0.32	0.09	0.00
	16	0.19	0.06	0.00	0.19	0.06	0.00	0.19	0.06	0.00
	23	0.33	0.21	0.11	-0.31	0.17	0.07	-0.47	0.21	0.02

SE: standard error

In the UK-USA sample, on the other hand, a statistically significant and uniform DIF with a small effect size was detected in the direction of the UK in item 10 and in the direction of the USA in item 16. In item 23, however, the non-uniform DIF effect, which is statistically significant only in HAC, is moderately large in the direction of the UK.

In Table 4, DIF effects with an effect size below 0.45 are shown as A, above 0.64 are shown as C, and between these two values are shown as B. Additionally, DIF effects in favor of the UK (reference group) are shown with “-” and in favor of the other country (focal group) “+”.

Table 4. *Direction and magnitude of DIF effects.*

UK-Türkiye				UK-Finland				UK-Japan				UK-USA			
Item	LAC	MAC	HAC	Item	LAC	MAC	HAC	Item	LAC	MAC	HAC	Item	LAC	MAC	HAC
5	A+*	A+*	A+*	1	A+	A+	A+	1	A+	A+	A+	10	A-	A-	A-
7	A+	A+	A+	7	A+	A+	A+	4	A-	A-	A-	16	A+	A+	A+
11	A+	A+	A+	8	A+	A+	A+	8	A+*	A-	A-	23	A+*	A-*	B-
13	A-	B-	A-*	11	A+*	C+*	C+*	9	B-	B-	B-				
14	A+	A+	A+	13	B-	A-	A+*	11	A+	A+	A+				
15	A-*	A-	A-*	15	A+	A+	C-*	12	A-*	A-	C-*				
16	A-*	A-	C-*	19	A-	A-	A-	13	A-	A-	A-				
20	A-	A+*	A+*	22	A-	A-	A-	15	A+	A+	C-*				
21	A-	A-	A-					17	A+	A+	A+				
22	A-	A-	A-					20	B-	B-	B-				
								23	A-	A-	A-				

*$p>0.05$

According to this information, the uniform DIF coefficients can be interpreted as follows. In all latent classes, the probability of answering item 11 correctly for students in the Türkiye sample is approximately 1.46 times that of students in the UK sample ($e^{0.38} = 1.46$). The probability of students in the Finland sample answering item 7 correctly is approximately 1.30 times the probability of answering correctly for students in the UK sample ($e^{0.26} = 1.30$). The probability of students in the UK sample answering item 9 correctly is approximately 1.68 times that of students in the Japan sample ($e^{0.52} = 1.68$). The probability of students in the UK sample answering item 10 correctly is approximately 1.38 times the probability of answering item 10 correctly than the students in the US sample ($e^{0.32} = 1.38$).

However, according to non-uniform DIF coefficients, the probability of answering item 13 correctly for students in the UK sample in MAC is approximately 1.58 times the probability of answering correctly for students in the Türkiye sample ($e^{0.46} = 1.58$). In LAC, the probability of students in the UK sample answering item 13 correctly is approximately 1.70 times the probability of answering correctly for students in the Finland sample ($e^{0.53} = 1.70$). In MAC, the probability of students in the Japan sample answering item 15 correctly is approximately 1.35 times the probability of answering item 15 correctly than the students in the UK sample ($e^{0.30} = 1.35$). In HAC, the probability of students in the UK sample answering item 23 correctly is approximately 1.60 times the probability of answering item 23 correctly than the students in the US sample ($e^{0.47} = 1.60$). Other DIF effects can be interpreted similarly.

4. DISCUSSION and CONCLUSION

In this study, whether the PISA 2018 application mathematical literacy test items in booklet number three show DIF across countries was examined with the latent class MIMIC approach. The UK was chosen as the reference group, and Türkiye, Finland, Japan and the USA as the focal group.

Considering the number of items with DIF detected according to the country variable in the paired comparisons examined, it was seen that fewer items showed DIF in the UK-USA sample (three items) compared to other samples (UK-Türkiye nine items, UK-Finland seven items, UK-Japan 11 items). There is one item with a statistically significant B level DIF in the UK-Türkiye sample, one in the UK-Finland sample, two in the UK-Japan sample, and one item in the UK-USA sample. No statistically significant C level DIF effect was observed in any of the samples caused by the country variable. The fact that the number of items with DIF observed in the UK-USA sample and their effect sizes are considerably less than in other samples strengthen the opinion that the source of DIF in other samples is significantly related to test language. In addition, more DIF items were observed in the UK-Japan sample in terms of number and effect size compared to other samples. This again showed the importance of translation between languages and differences between cultures in adaptation studies. However, the items should be analyzed qualitatively to determine whether the DIF in the related items is due to the real difference between the groups or bias.

In the four sample examinations, different items showed DIF in different samples. However, it was also observed that some items showed DIF in the same direction in both samples. For example, in both the UK-Türkiye and the UK-Finland samples, item 7 showed DIF at level A in favor of the focal group, and item 22 showed DIF at level A in favor of the reference group. A similar situation can be said for items 1 and 15 in the UK-Finland and the UK-Japan samples. In addition, only item 13 showed DIF in favor of the reference group in the other three samples except the UK-USA, but with different effect sizes in different latent classes. Item 13 shows DIF at level A for all latent classes in the UK-Japan sample. But in the UK-Türkiye sample, level A DIF for LAC and level B DIF for MAC; and in the UK-Finland sample level B DIF for LAC and level A DIF for MAC was observed. Similarly, Saatçioğlu (2022) examined the DIF of PISA 2018 financial literacy items resulting from the gender variable using the latent class MIMIC method. As a result, it was determined that the DIF effect differed (non-uniformly) in latent classes in 5 out of 16 test items.

As mentioned before and as seen in this study, the LCA approach allows the examination of test items in terms of DIF not only according to the observed variables but also for the latent classes. As Zumbo et al. (2015) and Elkonca (2020) stated, it is thought that this will enable the DIF sources to be determined in more detail and accurately. However, it is seen that the effect and size of DIF in non-homogeneous groups differ between groups, and these effects can be examined in more detail with the LCA method. This is in line with the results of Oliveri et al. (2016), Sawatzky et al. (2018), and Uyar (2020).

4.1. Suggestions

1. At the end of the analysis, it was seen that some of the items whose DIF effect was detected in the second and fourth steps were not statistically significant in the final model (M5.0) (item 5 in the UK-Türkiye sample and item 11 in the UK-Finland sample). In future studies, as Masyn (2017) and Tsaousis et al. (2020) suggested, sequential procedures according to p values in DIF determination steps or simultaneous procedures in terms of DIF type can be tried in terms of reviewing and improving the latent class MIMIC procedures used in this research, and simulation and real data studies can be done to investigate Type I and Type II errors.

2. Different DIF determination methods can be compared with the method used in the research and the conditions under which the methods are strong or weak relative to each other can be investigated.

3. In this research, we examined only mathematical literacy test in PISA 2018 and only booklet number three. Other tests or booklets in the application can be examined in terms of different observed variables (such as gender, region of residence of the student, and socioeconomic structure).

4. Test developers should better consider the characteristics of countries, such as curriculum, language, and culture, in both test development and adaptation studies and should do their part more carefully to avoid situations that may cause bias.

Acknowledgments

This study has been produced from Master's Thesis that was conducted under the supervision of the Prof. Dr. Tuncay Öğretmen and prepared by Selim Daşçıoğlu.

Declaration of Conflicting Interests and Ethics

The authors declare 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 authors.

Contribution of Authors

Selim Daşçıoğlu: Investigation, Visualization, Conception, Methodology, Analysis, and Writing-original draft. **Tuncay Öğretmen:** Conception, Supervision, Software, Critical Review, and Validation.

Orcid

Selim Daşçıoğlu https://orcid.org/0000-0001-6820-4585
Tuncay Öğretmen https://orcid.org/0000-0001-7783-1409

REFERENCES

Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items*. SAGE Publications.

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.

Elkonca, F. (2020). *ABİDE öz yeterlilik ölçeği DMF kaynaklarının gizil sınıf yaklaşımıyla incelenmesi [An analysis of the DIF sources of ABİDE self-efficacy scale by means of a latent class approach]* [Unpublished doctoral dissertation]. Gazi University.

Fraenkel, J.R., Wallen, N.E., & Hyun, H.H. (2011). *How to design and evaluate research in education*. McGraw-Hill Education.

Güngör Culha, D. (2012). *Örtük sınıf analizlerinde ölçme eşdeğerliğinin incelenmesi [Investigating measurement equivalence with latent class analysis]* [Unpublished doctoral dissertation]. Ege University.

Hambleton, R.K., Merenda, P.F., & Spielberger, C.D. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Lawrence Erlbaum Associates.

Harrington, D. (2009). *Confirmatory factor analysis*. Oxford University Press.

Kerlinger, F.N. (1999). *Foundations of behavioral research*. Wadsworth Publishing.

Lanza, S., & Collins, L. (2009). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. John Wiley & Sons, Inc.

Magidson, J., & Vermunt, J.K. (2004). Latent class models. D. Kaplan (Eds.), *The sage handbook of quantitative methodology for the social sciences* (s. 175-198). Sage Publications.

Masyn, K.E. (2017). Measurement invariance and differential item functioning in latent class analysis with stepwise multiple indicator multiple cause modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(2) 180-197.

McCutcheon, A.L. (1987). *Latent class analysis.* Sage Publication.

MEB (2019). *PISA 2018 Türkiye ön raporu [PISA 2018 Results]*. T.C. Milli Eğitim Bakanlığı.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Research Article, 18*(2), 5-11.

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory third edition*. McGraw-Hill.

Nylund-Gibson, K., Grimm, R., Quirk, M., & Furlong, M. (2014). A latent transition mixture model using the three-step specification. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(3), 439-454.

OECD. (2016a). *Sampling in PISA*. OECD Publishing.

OECD. (2016b). *PISA 2018 technical report*. OECD Publishing.

OECD. (2016c). *PISA 2018 translation and adaptation guidelines*. OECD Publishing.

OECD. (2019). PISA 2018 mathematics framework. *PISA 2018 assessment and analytical framework* (s. 73-95). OECD Publishing. https://doi.org/10.1787/13c8a22c-en

Oliveri, M.E., Ercikan, K., Lyons-Thomas, J., & Holtzman, S. (2016). Analyzing fairness among linguistic minority populations using a latent class differential item functioning approach. *Applied Measurement in Education, 29*(1), 17-29. https://doi.org/10.1080/08957347.2015.1102913

Saatçioğlu, F.M. (2022). Differential item functioning across gender with MIMIC modeling: PISA 2018 financial literacy items. *International Journal of Assessment Tools in Education*, *9*(3), 631-653. https://doi.org/10.21449/ijate.1076464

Sawatzky, R., Russell, L.B., Sajobi, T.T., Lix, L.M., Kopec, J., & Zumbo, B.D. (2018). The use of latent mixture models to identify Invariant Items in test construction. *Qual Life Res, 27*(7), 1745-1755. https://doi.org/10.1007/s11136-017-1680-8

Tsaousis, I., Sideridis, G.D., & AlGhamdi, H.M. (2020). Measurement invariance and differential item 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*(622). https://doi.org/10.3389/fpsyg.2020.00622

Uyar, Ş. (2020). Latent class approach to detect differential item functioning: PISA 2015. *Eurasian Journal of Educational Research*, *20*(88), 179-198. https://doi.org/10.14689

Vermunt, J.K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*(4), 450-469.

Zumbo, B.D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): logistic regression modeling as a unitary framework for binary and Likert-Type (Ordinal) item scores*. ON: Directorate of Human Resources Research and Evaluation.

Zumbo, B.D., Liu, Y., Wu, A.D., Shear, B.R., Olvera Astivia, O.L., & Ark, T.K. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly, 12*(1), 136-151. https://doi.org/10.1080/15434303.2014.972559

APPENDIX

Table 5. *Latent class MIMIC analysis results in the UK-Türkiye sample.*

Step	Model	LL	Npar	Comparison	Δ-2LL	ΔNpar	*p*
1	M1.0	-7979.74	73	M1.0 - M1.1	263.21	69	0.00
	M1.1	-7848.13	142				
2	M2.0.1	-1844,19	7	M2.0.1 - M2.1.1	3.68	3	0.30
	M2.1.1	-1842.35	10				
	M2.0.2	-1832.64	7	M2.0.2 - M2.1.2	1.73	3	0.63
	M2.1.2	-1831.77	10				
	M2.0.3	-1779.28	7	M2.0.3 - M2.1.3	6.08	3	0.11
	M2.1.3	-1776.24	10				
	M2.0.4	-1825.93	7	M2.0.4 - M2.1.4	1.57	3	0.67
	M2.1.4	-1825,14	10				
	M2.0.5	-1737.37	7	M2.0.5 - M2.1.5	12.30	3	0.01
	M2.1.5	-1731.23	10				
	M2.0.6	-1526.47	7	M2.0.6 - M2.1.6	0.90	3	0.83
	M2.1.6	-1526.01	10				
	M2.0.7	-1851.96	7	M2.0.7 - M2.1.7	8.44	3	0.04
	M2.1.7	-1847.74	10				
	M2.0.8	-1835.26	7	M2.0.8 - M2.1.8	3.90	3	0.27
	M2.1.8	-1833.31	10				
	M2.0.9	-1769.20	7	M2.0.9 - M2.1.9	7.42	3	0.06
	M2.1.9	-1765.49	10				
	M2.0.10	-1640.81	7	M2.0.10 - M2.1.10	1.00	3	0.80
	M2.1.10	-1640.31	10				
	M2.0.11	-1645.25	7	M2.0.11 - M2.1.11	27.90	3	0.00
	M2.1.11	-1631.30	10				
	M2.0.12	-1666.29	7	M2.0.12 - M2.1.12	1.46	3	0.69
	M2.1.12	-1665.56	10				
	M2.0.13	-1822.90	7	M2.0.13 - M2.1.13	35.71	3	0.00
	M2.1.13	-1805.04	10				
	M2.0.14	-1827,19	7	M2.0.14 - M2.1.14	11.81	3	0.01
	M2.1.14	-1821.29	10				
	M2.0.15	-1765.14	7	M2.0.15 - M2.1.15	13.38	3	0.00
	M2.1.15	-1758.45	10				
	M2.0.16	-1786.84	7	M2.0.16 - M2.1.16	16.09	3	0.00
	M2.1.16	-1778.79	10				
	M2.0.17	-1598.18	7	M2.0.17 - M2.1.17	6.03	3	0.11
	M2.1.17	-1595.17	10				
	M2.0.18	-1501.41	7	M2.0.18 - M2.1.18	1.68	3	0.64
	M2.1.18	-1500.57	10				
	M2.0.19	-1825.37	7	M2.0.19 - M2.1.19	3.01	3	0.39
	M2.1.19	-1823.87	10				
	M2.0.20	-1754.93	7	M2.0.20 - M2.1.20	8.37	3	0.04

Table 5. *(Continued)*

Step	Model	LL	Npar	Comparison	Δ-2LL	ΔNpar	*p*
2	M2.1.20	-1750.75	10				
	M2.0.21	-1789.39	7	M2.0.21 - M2.1.21	13.62	3	0.00
	M2.1.21	-1782.58	10				
	M2.0.22	-1707.47	7	M2.0.22 - M2.1.22	11.99	3	0.01
	M2.1.22	-1701.47	10				
	M2.0.23	-1610.24	7	M2.0.23 - M2.1.23	3.65	3	0.30
	M2.1.23	-1608.42	10				
3	M3.0	-7883.71	103	M1.0 - M3.0	192.05	30	0.00
				M3.0 - M1.1	71.16	39	0.00
4	M4.1	-7883.94	101	M4.1 - M3.0	0.44	2	0.80
	M4.2	-7883.73	101	M4.2 - M3.0	0.04	2	0.98
	M4.3	-7884.49	101	M4.3 - M3.0	1.55	2	0.46
	M4.4	-7888.26	101	M4.4 - M3.0	9.10	2	0.01
	M4.5	-7883.72	101	M4.5 - M3.0	0.02	2	0.99
	M4.6	-7886.75	101	M4.6 - M3.0	6.08	2	0.05
	M4.7	-7888.33	101	M4.7 - M3.0	9.22	2	0.01
	M4.8	-7887.52	101	M4.8 - M3.0	7.61	2	0.02
	M4.9	-7884.43	101	M4.9 - M3.0	1.44	2	0.49
	M4.10	-7885.33	101	M4.10 - M3.0	3.24	2	0.20
5	M5.0	-7887,18	91	M5.0 - M3.0	6.93	12	0.86

Table 6. *Latent class MIMIC analysis results in the UK-Finland sample.*

Step	Model	LL	Npar	Comparison	Δ-2LL	ΔNpar	*p*
1	M1.0	-7268.68	73	M1.0 - M1.1	246.06	69	0.00
	M1.1	-7145.65	142				
2	M2.0.1	-1633.39	7	M2.0.1 - M2.1.1	14.16	3	0.00
	M2.1.1	-1626.31	10				
	M2.0.2	-1680.58	7	M2.0.2 - M2.1.2	5.44	3	0.14
	M2.1.2	-1677.86	10				
	M2.0.3	-1634.01	7	M2.0.3 - M2.1.3	1.15	3	0.76
	M2.1.3	-1633.43	10				
	M2.0.4	-1674.09	7	M2.0.4 - M2.1.4	5.69	3	0.13
	M2.1.4	-1671.25	10				
	M2.0.5	-1566.33	7	M2.0.5 - M2.1.5	1.41	3	0.70
	M2.1.5	-1565.62	10				
	M2.0.6	-1383,12	7	M2.0.6 - M2.1.6	0.32	3	0.96
	M2.1.6	-1382.95	10				
	M2.0.7	-1667.79	7	M2.0.7 - M2.1.7	26.40	3	0.00
	M2.1.7	-1654.59	10				
	M2.0.8	-1635.72	7	M2.0.8 - M2.1.8	9.04	3	0.03
	M2.1.8	-1631.20	10				
	M2.0.9	-1585.90	7	M2.0.9 - M2.1.9	2.49	3	0.48

Table 6. *(Continued)*

Step	Model	LL	Npar	Comparison	Δ-2LL	ΔNpar	*p*
	M2.1.9	-1584.66	10				
	M2.0.10	-1507.85	7	M2.0.10 - M2.1.10	0.20	3	0.98
	M2.1.10	-1507.75	10				
	M2.0.11	-1495.75	7	M2.0.11 - M2.1.11	16.21	3	0.00
	M2.1.11	-1487.65	10				
	M2.0.12	-1498.40	7	M2.0.12 - M2.1.12	5.03	3	0.17
	M2.1.12	-1495.88	10				
	M2.0.13	-1665.78	7	M2.0.13 - M2.1.13	50.38	3	0.00
	M2.1.13	-1640.58	10				
	M2.0.14	-1634.23	7	M2.0.14 - M2.1.14	7.39	3	0.06
	M2.1.14	-1630.53	10				
	M2.0.15	-1613.60	7	M2.0.15 - M2.1.15	15.83	3	0.00
	M2.1.15	-1605.68	10				
	M2.0.16	-1579.34	7	M2.0.16 - M2.1.16	1.23	3	0.75
	M2.1.16	-1578.72	10				
	M2.0.17	-1454.75	7	M2.0.17 - M2.1.17	0.54	3	0.91
	M2.1.17	-1454.48	10				
	M2.0.18	-1349.08	7	M2.0.18 - M2.1.18	0.01	3	1.00
	M2.1.18	-1349.07	10				
2	M2.0.19	-1665.61	7	M2.0.19 - M2.1.19	14.83	3	0.00
	M2.1.19	-1658.20	10				
	M2.0.20	-1588.13	7	M2.0.20 - M2.1.20	0.54	3	0.91
	M2.1.20	-1587.86	10				
	M2.0.21	-1615.46	7	M2.0.21 - M2.1.21	2.09	3	0.55
	M2.1.21	-1614.41	10				
	M2.0.22	-1581.46	7	M2.0.22 - M2.1.22	19.09	3	0.00
	M2.1.22	-1571.91	10				
	M2.0.23	-1479.37	7	M2.0.23 - M2.1.23	2.87	3	0.41
	M2.1.23	-1477.93	10				
3	M3.0	-7175.34	97	M1.0 - M3.0	186.68	24	0.00
				M3.0 - M1.1	59.37	45	0.07
4	M4.1	-7177.62	95	M4.1 - M3.0	4.58	2	0.10
	M4.2	-7176.73	95	M4.2 - M3.0	2.79	2	0.25
	M4.3	-7175.50	95	M4.3 - M3.0	0.32	2	0.85
	M4.4	-7178.88	95	M4.4 - M3.0	7.10	2	0.03
	M4.5	-7179.90	95	M4.5 - M3.0	9.14	2	0.01
	M4.6	-7179.88	95	M4.6 - M3.0	9.09	2	0.01
	M4.7	-7177.54	95	M4.7 - M3.0	4.40	2	0.11
	M4.8	-7176.95	95	M4.8 - M3.0	3.24	2	0.20
5	M5.0	-7182.77	87	M5.0 - M3.0	14.86	10	0.14

Table 7. *Latent class MIMIC analysis results in the UK-Japan sample.*

Step	Model	LL	Npar	Comparison	Δ-2LL	ΔNpar	*p*
1	M1.0	-7881.14	73	M1.0 - M1.1	428.34	69	0.00
	M1.1	-7666.97	142				
2	M2.0.1	-1763.18	7	M2.0.1 - M2.1.1	13.77	3	0.00
	M2.1.1	-1756.29	10				
	M2.0.2	-1801.82	7	M2.0.2 - M2.1.2	3.88	3	0.27
	M2.1.2	-1799.88	10				
	M2.0.3	-1757.99	7	M2.0.3 - M2.1.3	2.69	3	0.44
	M2.1.3	-1756.65	10				
	M2.0.4	-1807.98	7	M2.0.4 - M2.1.4	38.34	3	0.00
	M2.1.4	-1788.81	10				
	M2.0.5	-1684.26	7	M2.0.5 - M2.1.5	6.05	3	0.11
	M2.1.5	-1681.23	10				
	M2.0.6	-1511.89	7	M2.0.6 - M2.1.6	0.92	3	0.82
	M2.1.6	-1511.44	10				
	M2.0.7	-1786.41	7	M2.0.7 - M2.1.7	4.54	3	0.21
	M2.1.7	-1784.14	10				
	M2.0.8	-1820,20	7	M2.0.8 - M2.1.8	29.20	3	0.00
	M2.1.8	-1805.60	10				
	M2.0.9	-1712.57	7	M2.0.9 - M2.1.9	69.78	3	0.00
	M2.1.9	-1677.68	10				
	M2.0.10	-1625.72	7	M2.0.10 - M2.1.10	0.44	3	0.93
	M2.1.10	-1625.49	10				
	M2.0.11	-1606.48	7	M2.0.11 - M2.1.11	19.54	3	0.00
	M2.1.11	-1596.71	10				
	M2.0.12	-1648.43	7	M2.0.12 - M2.1.12	17.28	3	0.00
	M2.1.12	-1639.79	10				
	M2.0.13	-1812.92	7	M2.0.13 - M2.1.13	8.26	3	0.04
	M2.1.13	-1808.79	10				
	M2.0.14	-1781.11	7	M2.0.14 - M2.1.14	3,58	3	0.31
	M2.1.14	-1779.32	10				
	M2.0.15	-1734.47	7	M2.0.15 - M2.1.15	25.06	3	0.00
	M2.1.15	-1721.94	10				
	M2.0.16	-1686.26	7	M2.0.16 - M2.1.16	6.49	3	0.09
	M2.1.16	-1683.02	10				
	M2.0.17	-1619.53	7	M2.0.17 - M2.1.17	23.48	3	0.00
	M2.1.17	-1607.79	10				
	M2.0.18	-1477.03	7	M2.0.18 - M2.1.18	0.51	3	0.92
	M2.1.18	-1476.78	10				
	M2.0.19	-1794.78	7	M2.0.19 - M2.1.19	6.84	3	0.08
	M2.1.19	-1791.36	10				
	M2.0.20	-1773.69	7	M2.0.20 - M2.1.20	85.47	3	0.00
	M2.1.20	-1730.96	10				
	M2.0.21	-1743.57	7	M2.0.21 - M2.1.21	6.29	3	0.10

Table 7. *(Continued)*

Step	Model	LL	Npar	Comparison	Δ-2LL	ΔNpar	*p*
2	M2.1.21	-1740.43	10				
	M2.0.22	-1705.71	7	M2.0.22 - M2.1.22	4.89	3	0.18
	M2.1.22	-1703.26	10				
	M2.0.23	-1600,48	7	M2.0.23 - M2.1.23	9.71	3	0.02
	M2.1.23	-1595.62	10				
3	M3.0	-7696.40	106	M1.0 - M3.0	369.47	33	0.00
				M3.0 - M1.1	58.87	36	0.01
4	M4.1	-7697.10	104	M4.1 - M3.0	1.39	2	0.50
	M4.2	-7698.16	104	M4.2 - M3.0	3,51	2	0.17
	M4.3	-7706.11	104	M4.3 - M3.0	19.41	2	0.00
	M4.4	-7696.54	104	M4.4 - M3.0	0.28	2	0.87
	M4.5	-7698.47	104	M4.5 - M3.0	4.13	2	0.13
	M4.6	-7699.63	104	M4.6 - M3.0	6.46	2	0.04
	M4.7	-7699.26	104	M4.7 - M3.0	5.72	2	0.06
	M4.8	-7702.96	104	M4.8 - M3.0	13.12	2	0.00
	M4.9	-7696.60	104	M4.9 - M3.0	0.38	2	0.83
	M4.10	-7696.97	104	M4.10 - M3.0	1.13	2	0.57
	M4.11	-7699,381	104	M4.11 - M3.0	5.95	2	0.05
5	M5.0	-7708.26	90	M5.0 - M3.0	23.70	16	0.10

Table 8. *Latent class MIMIC analysis results in the UK-USA sample.*

Step	Model	LL	Npar	Comparison	Δ-2LL	ΔNpar	*p*
1	M1.0	-7002.28	73	M1.0 - M1.1	108.35	69	0.00
	M1.1	-6948.11	142				
2	M2.0.1	-1548.69	7	M2.0.1 - M2.1.1	3.94	3	0.27
	M2.1.1	-1546.72	10				
	M2.0.2	-1556.39	7	M2.0.2 - M2.1.2	2.71	3	0.44
	M2.1.2	-1555.04	10				
	M2.0.3	-1505.45	7	M2.0.3 - M2.1.3	2.88	3	0.41
	M2.1.3	-1504.01	10				
	M2.0.4	-1538,00	7	M2.0.4 - M2.1.4	3,57	3	0.31
	M2.1.4	-1536.21	10				
	M2.0.5	-1452.23	7	M2.0.5 - M2.1.5	3.11	3	0.37
	M2.1.5	-1450.68	10				
	M2.0.6	-1272.42	7	M2.0.6 - M2.1.6	0.14	3	0.99
	M2.1.6	-1272.35	10				
	M2.0.7	-1561.46	7	M2.0.7 - M2.1.7	2.48	3	0.48
	M2.1.7	-1560.22	10				
	M2.0.8	-1534.03	7	M2.0.8 - M2.1.8	1.58	3	0.66
	M2.1.8	-1533.24	10				
	M2.0.9	-1474.56	7	M2.0.9 - M2.1.9	2.99	3	0.39
	M2.1.9	-1473.07	10				

Table 8. *(Continued)*

Step	Model	LL	Npar	Comparison	Δ-2LL	ΔNpar	*p*
2	M2.0.10	-1382.50	7	M2.0.10 - M2.1.10	10.68	3	0.01
	M2.1.10	-1377.16	10				
	M2.0.11	-1431.26	7	M2.0.11 - M2.1.11	0.62	3	0.89
	M2.1.11	-1430.95	10				
	M2.0.12	-1382.67	7	M2.0.12 - M2.1.12	1.55	3	0.67
	M2.1.12	-1381.90	10				
	M2.0.13	-1575.71	7	M2.0.13 - M2.1.13	3,54	3	0.32
	M2.1.13	-1573.94	10				
	M2.0.14	-1536.97	7	M2.0.14 - M2.1.14	3.61	3	0.31
	M2.1.14	-1535.16	10				
	M2.0.15	-1477.89	7	M2.0.15 - M2.1.15	2.15	3	0.54
	M2.1.15	-1476.82	10				
	M2.0.16	-1476.56	7	M2.0.16 - M2.1.16	11.40	3	0.01
	M2.1.16	-1470.86	10				
	M2.0.17	-1345.95	7	M2.0.17 - M2.1.17	1.96	3	0.58
	M2.1.17	-1344.98	10				
	M2.0.18	-1249.37	7	M2.0.18 - M2.1.18	2.74	3	0.43
	M2.1.18	-1248,00	10				
	M2.0.19	-1557.72	7	M2.0.19 - M2.1.19	1.70	3	0.64
	M2.1.19	-1556.87	10				
	M2.0.20	-1501.45	7	M2.0.20 - M2.1.20	2.93	3	0.40
	M2.1.20	-1499.98	10				
	M2.0.21	-1516.76	7	M2.0.21 - M2.1.21	2.25	3	0.52
	M2.1.21	-1515.64	10				
	M2.0.22	-1469.22	7	M2.0.22 - M2.1.22	2.81	3	0.42
	M2.1.22	-1467.82	10				
	M2.0.23	-1348.14	7	M2.0.23 - M2.1.23	8.27	3	0.04
	M2.1.23	-1344.01	10				
3	M3.0	-6982.99	82	M1.0 - M3.0	38.59	9	0.00
				M3.0 - M1.1	69.76	60	0.18
4	M4.1	-6983.53	80	M4.1 - M3.0	1.08	2	0.58
	M4.2	-6983.54	80	M4.2 - M3.0	1.11	2	0.57
	M4.3	-6987.32	80	M4.3 - M3.0	8.66	2	0.01
5	M5.0	-6984.10	78	M5.0 - M3.0	2.21	4	0.70