ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2025; 16(1); 13-29

Latent Class Analysis and DIF Testing in Mathematics Achievement: A Comparative Study of Korea and Türkiye Using MIMIC Modeling

Hyo Seob SONG* Hee Sun JUNG**

Abstract

This study examines the latent classes of mathematics achievement and investigates differential item functioning (DIF) between Korea and Türkiye. Moreover, it explores the influence of the country on the latent classes of mathematics achievement. To achieve this, data from eighth-grade students in TIMSS 2019 were analyzed using Latent Class MIMIC Modeling. The findings uncovered diverse latent classes of math achievement and detected both uniform and Non-uniform DIF between Korea and Türkiye. Furthermore, the country was found to significantly affect the latent class membership of math achievement. This study highlights the necessity of verifying the measurement invariance of indicator variables in latent class analysis (LCA). It also sheds light on areas where students performed favorably or unfavorably in mathematics achievement tests across these countries by investigating DIF. These findings have important implications for mathematics education in Korea and Türkiye.

Keywords: Mathematics Achievement; Latent Class Analysis (LCA); Multiple Indicator Multiple Cause (MIMIC) Modeling; Measurement Invariance; Differential Item Functioning (DIF); TIMSS 2019

Introduction

Mathematics significantly influences students' academic success and future career prospects (Guhl, 2019; Lubinski et al., 2014). Researchers in mathematics education have utilized international comparative studies (e.g., TIMSS, PISA) to evaluate students' academic achievement (Arıcan et al., 2016; Badri, 2019; Wang et al., 2023; Wiberg, 2019). Since its inception in 1995, the Trends in Mathematics and Science Study (TIMSS) has played a crucial role in assessing national-level mathematics achievement by comparing the relative performance of participating countries over time. Participating countries use assessment results to improve their educational curricula and methods or to enhance achievement (Lee & Stankov, 2018; Şen & Arıcan, 2015). Additionally, TIMSS promotes efforts to advance STEM (Science, Technology, Engineering, Mathematics) education by providing participating countries with data on students' mathematics and science achievement levels (Geesa et al., 2020; Mullis & Martin, 2017). According to the results of TIMSS 2019 conducted by the IEA, there were differences in mathematics achievement among participating countries. Korea achieved a high level of achievement in mathematics, ranking among the top performers, while Türkiye recorded achievement around the international average (Mullis et al., 2020). Such differences in mathematics achievement among countries may arise from students' home resources, attitudes toward mathematics,

* Assistant Professor, U1 University, Institute for Educational Innovation, Chung cheong do - South Korea, songh-s@hanmail.net, ORCID ID: 0000-0001-7554-2849
** Professor, Sungkyunkwan University, Department of Mathematics Education, Seoul - South Korea, hsun90@skku.edu, ORCID ID: 0000-0003-0093-2193

To cite this article:
Song, H. S. & Jung, H. S. (2025). Latent class analysis and dif testing in mathematics achievement: A comparative study of Korea and Türkiye using MIMIC modeling. *Journal of Measurement and Evaluation in Education and Psychology, 16*(1), 13-29. https://doi.org/10.21031/epod.1539828

Received: 28.08.2024
Accepted: 7.03.2025

and cultural differences (Geesa et al., 2019; Klieme & Baumert, 2001), as well as variations in educational curricula across countries (Sohn, 2010). Particularly interesting in the results of TIMSS 2019 between Korea and Türkiye is that while Korea's mathematics achievement was significantly higher than that of Türkiye, Turkish students showed higher mathematics attitudes related to affective achievement compared to Korean students (Mullis et al., 2020). Korea's high mathematics achievement can be attributed to its society's strong emphasis on education, competitive examination, and selection systems (Im & Park, 2010), as well as participation in additional extracurricular education beyond school classes (Dittrich & Neuhaus, 2023; Shin et al., 2019; Woo & Hodges, 2015). Also contributing to the high math achievement of Korean students is the high quality of public education (Im & Park, 2010; Şen & Arıcan, 2015), which includes the implementation of constructivist teaching methods (Hwang & Hwang, 2008) and the competence of math teachers (Ko & Jung, 2020).

Recently, finite mixture models such as Latent Class Analysis (LCA) have been utilized across various research fields, including behavioral science, education, and psychology. Generally, research that applies finite mixture models involves investigating the relationship between predictor and latent class membership (Masyn, 2017; Song et al., 2023; Vermunt, 2010). The integration of predictors and the results of latent class membership has been evolving, and discussions have been held in several studies regarding the timing and method of including predictor variables in mixture models (Masyn, 2017; Nylund-Gibson & Masyn, 2016). Particularly, the 3-step method in latent class modeling is known to produce more robust and accurate results compared to the 1-step method. This is because it excludes covariates in the step of class enumeration, thereby eliminating the risk of class composition varying depending on covariates. However, previous studies have reported that biased estimates of the effects of covariates on latent class variables may occur if the direct effects of covariates on indicator variables are ignored in the 3-step method (Asparouhov & Muthén, 2014; Masyn, 2017). This implies that to estimate the effects of covariates on latent class variables, it is necessary to conduct measurement invariance tests. These tests confirm the direct effects of covariates on each indicator variable within each latent class. This process follows the completion of class enumeration using an unconditional latent class model in the first step of the 3-step method. Based on previous studies that have shown ignoring the direct effects of covariates on indicator variables in LCA can lead to biased estimates of the effects of covariates on latent classes (Clark & Muthén, 2009; Nylund-Gibson & Choi, 2018), Masyn (2017) proposed a method for detecting these direct effects in LCA. Masyn's method combines LCA with the multiple indicator multiple cause (MIMIC) model to confirm the measurement invariance of indicator variables across covariates. This approach enables accurate estimation of the effects of covariates on latent classes and exploration of DIF of indicator variables by covariates. DIF in latent class MIMIC models refers to items where individuals belonging to the same latent class exhibit different expected responses depending on the values of covariates (Masyn, 2017). Uniform DIF is assessed when the difference in expected responses to indicators by covariates is consistent across all classes, while non-uniform DIF is assessed when the difference in expected responses to indicators by covariates varies across one or more classes (Masyn, 2017). Latent classes emerge when not all members exhibit homogeneous response patterns (De Ayala et al., 2002; Samuelsen, 2008). Particularly, results of exploring DIF obtained from the entire population may be biased, thus studies on DIF should be examined across latent classes (Saaatcioglu, 2022). In the studies by Tsaousis, Sideridis, AlGhamdi (2020) and Saaatcioglu (2022), the method proposed by Masyn (2017) was used to explore gender-specific DIF in achievement tests, investigating DIF by gender in the latent class of academic achievement.

To compare academic achievement among countries with different languages and cultures, scale measurement invariance must be secured first (Hambleton, 2001). Recently, a growing body of research has focused on assessing and exploring the causes of measurement invariance across different languages, cultures, and countries in international achievement tests (Demirus & Pektas, 2022; Im & Park, 2010; Sohn, 2010; Yoon & Lee, 2013). Most of these studies apply the technique of DIF to assess the level of equivalence at the item level. For example, Im and Park (2010) compared the mathematics scores of 8th-grade students in Korea and the United States using TIMSS 2003 data, revealing variations in problem reformulation, inference, measurement, and geometry. Demirus and Pektas (2022) examined the presence of DIF in the multiple-choice items of the TIMSS 2015 science achievement test across various countries, including Türkiye, Australia, New Zealand, Morocco, and Egypt. Their study confirmed that more instances of DIF were observed between countries with diverse cultures and languages, suggesting that language variations contributed to DIF. Sohn (2010) identified DIF between Korean and Finnish students using PISA 2006 mathematics test data. Yoon and Lee (2013) investigated DIF on the TIMSS 2007 mathematics test among students from Korea, the United States, and Singapore. International comparative research using DIF enables the identification of item characteristics that function differentially when compared across countries, even when individuals have similar abilities. This provides insights into the strengths and weaknesses of domestic students and serves as foundational data for improving educational curricula and environments (Sohn, 2010).

This study aims to explore the latent classes of mathematics achievement in the TIMSS 2019 mathematics assessment using the Latent Class MIMIC Modeling proposed by Masyn (2017). It focuses on 8th-grade students in two countries: Korea, the top-performing country on the TIMSS 2019 mathematics test, and Türkiye, which performs around the international average but has been steadily increasing its achievement since joining TIMSS. Additionally, this study explores differential item functioning (DIF) to verify measurement invariance in the mathematics achievement test between Korea and Türkiye. DIF occurs due to violations of measurement invariance across different subgroups (Huang, 2020). Furthermore, it investigates the influence of the country (Korea/Türkiye) on the latent class membership of mathematics achievement. The research questions of this study are as follows:

1. How are latent classes of mathematics achievement identified in combined Korean and Turkish students?

2. Does DIF exist in the mathematics achievement test between Korea and Türkiye?

3. Does the country (Korea/Türkiye) influence the latent class membership of mathematics achievement?

Methods

Data

In this research, data from 8th-grade students in South Korea and Türkiye who participated in TIMSS 2019 were examined. The Trends in International Mathematics and Science Study (TIMSS) is an international assessment of academic performance organized by the International Association for the Evaluation of Educational Achievement (IEA). This assessment measures students' mathematics and science achievements at a global level to evaluate and enhance educational outcomes (Mullis et al., 2020). Initiated in 1995, TIMSS is conducted every four years, targeting 4th-grade and 8th-grade students. The assessment includes mathematics and science achievement tests based on the curricula of the participating countries, along with surveys of schools, teachers, students, and parents about educational contextual factors (Mullis et al., 2020). The TIMSS 2019 8th-grade mathematics assessment comprises 211 items. The framework is divided into two dimensions: the content dimension (Number, Algebra, Geometry, Data and Probability) detailing the subject matter, and the cognitive dimension (Knowing, Applying, and Reasoning) outlining the thinking processes evaluated as students engage with the content (Mullis & Martin, 2017).

A final sample of 553 South Korean students and 582 Turkish students, who participated in Booklet 5 and 6 of the TIMSS 2019 8th-grade mathematics assessment, was selected for this study, as shown in Table 1. The analysis included items from Block 6 of Booklets 5 and 6. Item ME62342, which had missing data for all countries, was excluded from the analysis. Thus, a total of 14 items were analyzed.

Table 1

The number of cases for analysis

Booklet	Excluded Items	Optional Items	Excluded Items	Korea	Türkiye	Total
Booklet 5	Block5	Block6 (ME62150~ ME62123B)		282	290	572
Booklet 6			Block7	271	292	563
Total				553(48.7%)	582(51.3%)	1,135(100%)

Data Analysis

In this research, Latent Class MIMIC Modeling was applied to identify the latent classes of mathematics performance and to examine measurement invariance and DIF of mathematics test items between Korea and Türkiye. Before conducting the analysis, test items were coded with correct answers as 1 and incorrect answers as 0. The country variable was coded as 1 for Korea and 0 for Türkiye. To determine the best number of latent classes for mathematics achievement, various criteria such as information criteria, scree plots, and entropy indices were used, along with considerations for interpretability and discriminant validity between groups (Ram & Grimm, 2009). Likelihood ratio tests were utilized to compare latent class MIMIC models, with effect sizes of identified DIF items evaluated using the Educational Testing Service (ETS) criteria. According to ETS guidelines, a logit value below 0.43 suggests a negligible DIF effect, a value of 0.43 or higher indicates a moderate effect, and a value of 0.64 or higher points to a large effect (Dorans & Holland, 1992). The analysis was performed using Mplus (Version 8.3) and the MplusAutomation package in R (Version 4.2.2), adhering to the method proposed by Masyn (2017), with some modifications detailed as follows:

Step 0: Conduct LCA to identify the optimal number of latent classes. Covariates are included as auxiliary variables to ensure they do not affect the identification of latent classes.

Step 1: Compare a baseline model (M_1.0, No_DIF), where covariates affect latent classes but not indicator variables, with an alternative model (M_1.1, All_DIF), where covariates directly affect both latent classes and all indicator variables. Acceptance of the baseline model (M_1.0) indicates no DIF for individual indicators by covariates, while acceptance of the alternative model (M_1.1) suggests the presence of DIF items for individual indicators by covariates, indicating at least one DIF item in at least one latent class.

Step 2: Conduct an omnibus DIF test to examine DIF for each indicator variable by covariates. This involves comparing model M_2.0.X (covariates affect latent classes but not indicator variables) with model M_2.1.X (covariates have direct effects on both latent classes and indicator variables).

Step 3: Select the optimal model by comparing model M_3.0, where all identified DIF items are treated as non-uniform DIF, with the baseline model (M_1.0, No_DIF) and the alternative model (M_1.1, All_DIF).

Step 4: Determine if the items identified as DIF in Step 2 are uniform DIF items by comparing the fit of model M_4.X (imposes uniform constraints on covariate effects on indicator variables across classes) with model M_3.0 (treats all identified DIF items as non-uniform DIF). If the fit of M_4.X is not significantly worse than that of M_3.0, the item is considered a uniform DIF.

Step 5: Choose the optimal model by comparing the fit of model M_5.0 (covariate effects on indicator variables are equal across latent classes for all identified uniform DIF items) with model M_3.0 (treats all identified DIF items as non-uniform DIF).

Step 6: Select the final model by comparing model M_6.0 (regression coefficients of covariates on latent class membership are constrained to 0) with model M_6.1 (regression coefficients of covariates on latent class membership are freely estimated) in the model chosen from Step 5.

Results

Descriptive Statistics

When examining the item difficulty index for each item in both Korea and Türkiye, it was found that the item difficulty index for all items was higher in Korea compared to Türkiye. Specifically, as shown in Table 2, the item difficulty index for Korean students ranged from 0.41 to 0.92, whereas for Turkish students ranged from 0.09 to 0.60. Particularly, in item 11, the difference in item difficulty between the two countries was 0.57, indicating the largest discrepancy. Additionally, for items 2, 3, 6, 7, 9, and 14, the difference in item difficulty index between the two countries exceeded 0.3, highlighting a notable variation in item difficulty.

Table 2

Math 8th Block6 Item

No	Variable	Domain	Label	Item difficulty index	
				Korea	Türkiye
1	ME62150	Number/Knowing	“DIFFERENCE BETWEEN LOW TEMPERATURE IN CITY X AND Y”	0.79	0.50
2	ME62335	Number/Knowing	“SELECT EQUIVALENT RATIO TO 3:2”	0.92	0.60
3	ME62219	Number/Applying	“KATY ENLARGES A PHOTO - NEW HEIGHT”	0.74	0.38
4	ME62002	Number/Reasoning	“FILL IN BOXES TO MAKE THE SMALLEST PRODUCT”	0.48	0.31
5	ME62149	Algebra/Applying	“IDENTIFY EXPRESSION TO CALCULATE ROBIN'S EARNINGS”	0.48	0.35
6	ME62241	Algebra/Applying	“ROY'S PHONE BUSINESS - EQUATION FOR Y”	0.70	0.27
7	ME62105	Algebra/Reasoning	“AREA OF RECTANGLE WITH SIDES X AND 2X + 1”	0.65	0.27
8	ME62040	Geometry/Applying	“ESTIMATE AREA OF IRREGULAR SHAPE ON 1 CM GRID”	0.60	0.46
9	ME62288A	Geometry/Applying	“FIND VERTICES OF TRAPEZOIDS M AND N”	0.41	0.11
10	ME62288B	Geometry/Applying	“FIND VERTICES OF TRAPEZOIDS M AND N”	0.41	0.09

Table 2 (Continued)

Math 8th Block6 Item

No	Variable	Domain	Label	Item difficulty index	
				Korea	Türkiye
11	ME62173	Geometry/Reasoning	"FIND ANGLE X ON A FOLDED PIECE OF PAPER"	0.76	0.19
12	ME62133	Data and Probability/Applying	"BLACK AND WHITE MARBLES IN A BAG WITH REPLACEMENT"	0.70	0.54
13	ME62123A	Data and Probability/Knowing	"RELAY RACE - MEAN TIME OF RUNNERS "	0.81	0.59
14	ME62123B	Data and Probability/Applying	"RELAY RACE - MEAN TIME WHEN 2 RUNNERS IMPROVE"	0.72	0.36

Measurement Invariance and DIF

Step 0: Before verifying the measurement invariance of the indicator variables and exploring the presence of DIF according to covariates, it is essential to select the optimal number of latent classes. To achieve this, latent class analysis on mathematics achievement was conducted without including covariates, identifying latent classes among the combined Korean and Turkish students. The optimal number of latent classes was determined by comparing the model fit and simplicity indicators as presented in Table 3. As shown in Table 3, the best fit was observed when there were five latent classes. With a large sample size, the values of AIC and BIC tend to decrease as the number of groups increases, and the number of latent classes can be determined using a scree plot (Jedidi et al., 1997).

Examination of the scree plot in Figure 1 reveals that the values of most goodness-of-fit indicators decrease at a slower rate after three latent classes, and AWE shows an increase after three latent classes. Additionally, when there are three latent classes, the entropy index is 0.905, indicating good performance. After considering factors such as goodness-of-fit indices, statistical significance, discriminant between groups, presence of latent classes, and interpretability, the optimal number of latent classes was determined to be three.

Upon examining the composition of classified latent classes in Figure 2, Class 1 (284 participants, 25.0%) exhibited a generally high item difficulty index of over 0.7 for each item, indicating the highest level of mathematics achievement among the three latent classes. Class 2(377 participants, 33.2%) showed moderate levels of mathematics achievement among the three latent classes, with significant differences in item difficulty index for each item. Notably, the item difficulty index for Geometry/Applying items 9 and 10 were below 0.1. Class 3 (474 participants, 41.8%) exhibited an item difficulty index generally below 0.4 across all items, indicating the lowest level of mathematics achievement among the three latent classes.

Consequently, Class 1 to Class 3 were respectively named the high-achievement group, the moderate-achievement group, and the low-achievement group. The item difficulty index by latent class and country is shown in Figure 3, while Figure 4 illustrates the composition of each latent class by country (Korea and Türkiye).

Table 3

LCA Model Fit

Class	Par	LL	BIC	aBIC	CAIC	AWE	BLRT
1-Class	14	-10333	20764	20719	20778	20904	-
2-Class	29	-8364	16932	16840	16961	17223	<0.001
3-Class	44	-8009	16327	16188	16371	16769	<0.001
4-Class	59	-7935	16285	16097	16344	16877	<0.001
5-Class	74	-7878	16277	16042	16351	17020	<0.001

Note. “Par”=parameters, “LL”=log likelihood, “BIC”=bayesian information criterion, “aBIC”=sample size adjusted BIC, “CAIC”=consistent Akaike information criterion, “AWE”=approximate weight of evidence criterion, “BLRT”=bootstrapped likelihood ratio test p-value

Figure 1

Scree Plot

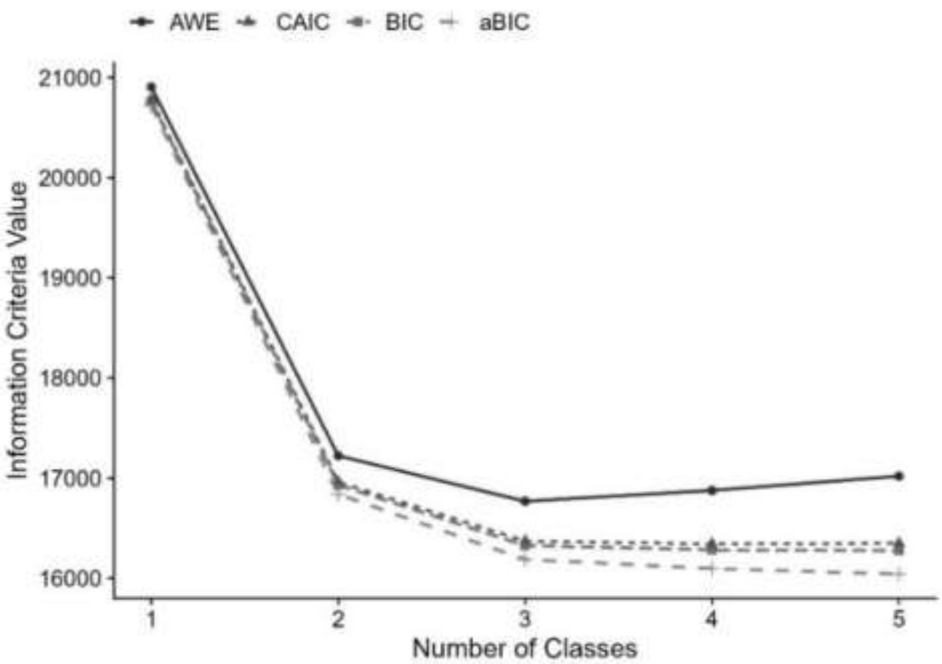


Figure 2

Latent class plots for Math Achievement

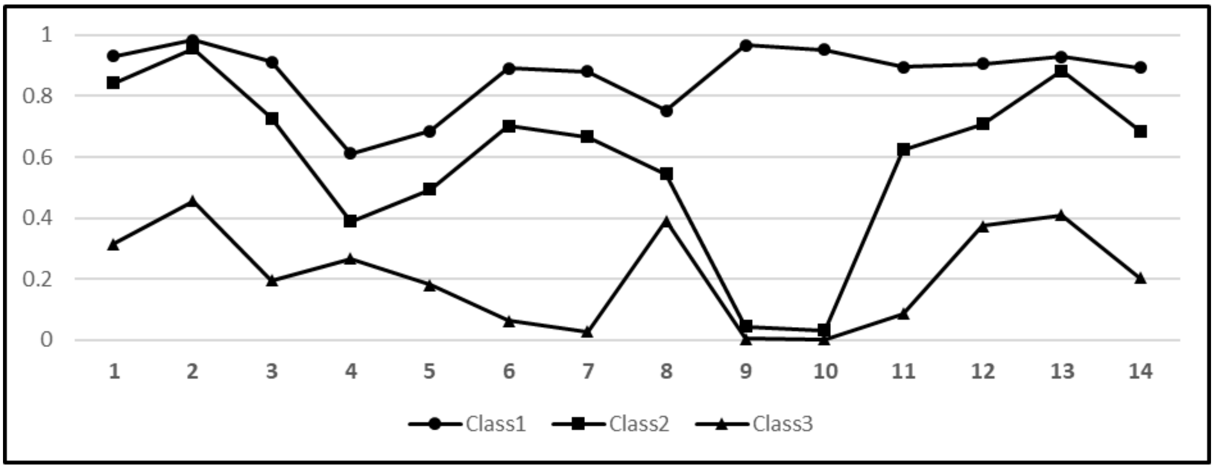


Figure 3

Item difficulty index within latent classes

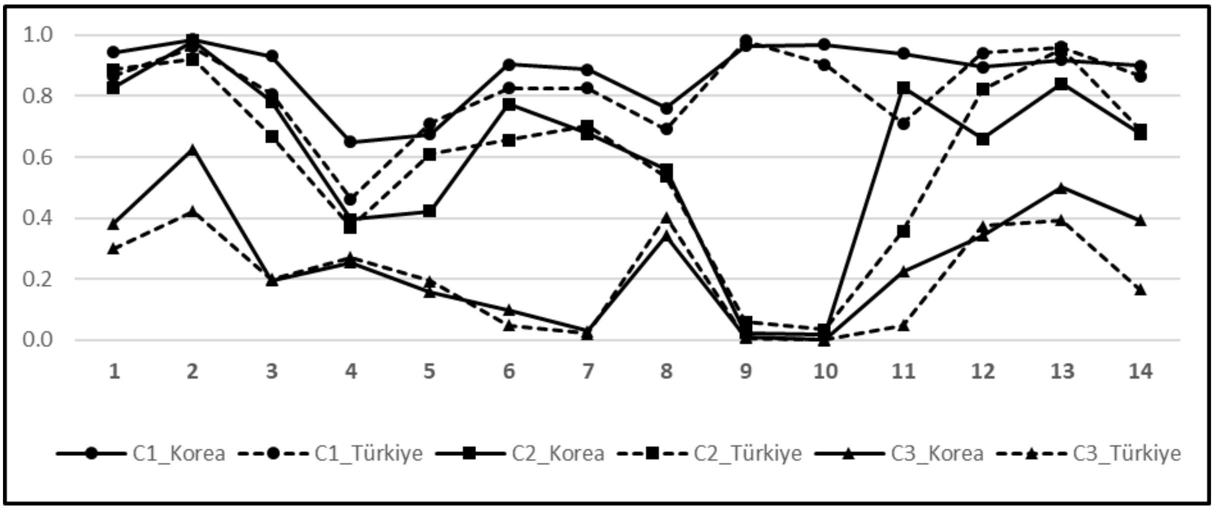
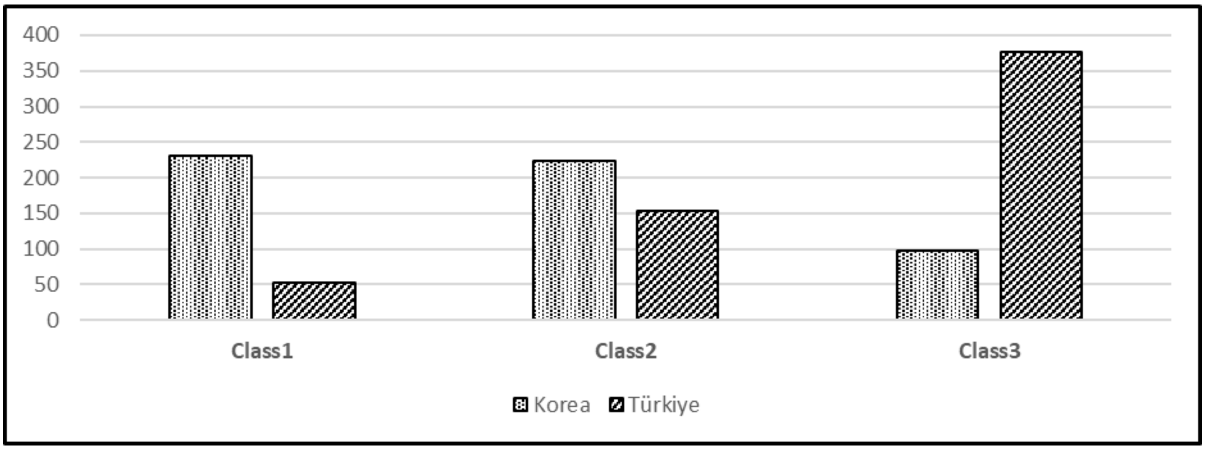


Figure 4

Composition of Korea & Türkiye within latent classes



Step 1: The latent class model selected in Step 0 was augmented with the covariate, the country variable, to compare the baseline model M_1.0 (No_DIF), where the country variable influenced the latent class variable but had no direct effects on the indicators, with the alternative model M_1.1 (All_DIF), where the country variable had direct effects on both the latent class variable and all indicators. As shown in Table 4, the fit of the M_1.1 model was significantly better than that of the M_1.0 model. This indicates that the country variable (Korea/Türkiye) is the source of DIF for at least one of the three latent classes and at least one of the fourteen items.

Step 2: DIF omnibus tests were conducted for each of the fourteen indicator variables by comparing models M_2.0.X, where the country variable (Korea/Türkiye) was set to influence the latent class variable but without direct effects on the indicator variables, and M_2.1.X, where the country variable was set to have direct effects on the indicator variables. As shown in Table 4, it was observed that for items 1, 4, 6, 7, 8, 9, and 10, the fit of the model without direct effects of the covariate on the indicators was not significantly worse than the model with direct effects. Additionally, for items 2, 3, 5, 11, 12, 13, and 14, the fit of the model with direct effects of the covariate on the indicators was significantly better than that without. This indicates that individually, seven items (2, 3, 5, 11, 12, 13, 14) out of the fourteen in the mathematics achievement test exhibit DIF.

Table 4

Model Comparisons for Stepwise DIF Test

Step	Model	Description	LL	npar	Comparison	LRTS	df	p
1	M_1.0	MIMIC: NO DIF	-7844.744	46	M_1.0 vs M_1.1	256.41	42	<0.001
	M_1.1	MIMIC: ALL DIF	-7716.539	88				
2	M_2.0.1	#1: No DIF	-1599.816	7	M_2.0.1 vs M_2.1.1	7.752	3	0.051
	M_2.1.1	#1: Non U DIF	-1595.940	10				
	M_2.0.2	#2: No DIF	-1480.067	7	M_2.0.2 vs M_2.1.2	10.871	3	0.012
	M_2.1.2	#2: Non U DIF	-1480.631	10				
	M_2.0.3	#3: No DIF	-1604.997	7	M_2.0.3 vs M_2.1.3	7.820	3	0.049
	M_2.1.3	#3: Non U DIF	-1601.087	10				
	M_2.0.4	#4: No DIF	-1784.847	7	M_2.0.4 vs M_2.1.4	6.978	3	0.072
	M_2.1.4	#4: Non U DIF	-1781.358	10				
	M_2.0.5	#5: No DIF	-1734.017	7	M_2.0.5 vs M_2.1.5	20.932	3	<0.001
	M_2.1.5	#5: Non U DIF	-1723.551	10				
	M_2.0.6	#6: No DIF	-1496.159	7	M_2.0.6 vs M_2.1.6	3.640	3	0.303
	M_2.1.6	#6: Non U DIF	-1494.339	10				
	M_2.0.7	#7: No DIF	-1464.368	7	M_2.0.7 vs M_2.1.7	4.588	3	0.205
	M_2.1.7	#7: Non U DIF	-1462.074	10				
	M_2.0.8	#8: No DIF	-1806.546	7	M_2.0.8 vs M_2.1.8	1.910	3	0.591
	M_2.1.8	#8: Non U DIF	-1805.591	10				
	M_2.0.9	#9: No DIF	-1184.171	7	M_2.0.9 vs M_2.1.9	5.773	3	0.123
	M_2.1.9	#9: Non U DIF	-1181.284	10				
	M_2.0.10	#10: No DIF	-1170.930	7	M_2.0.10 vs M_2.1.10	4.944	3	0.176
	M_2.1.10	#10: Non U DIF	-1168.466	10				
	M_2.0.11	#11: No DIF	-1517.144	7	M_2.0.11 vs M_2.1.11	91.378	3	<0.001

Table 4 (Continued)

Model Comparisons for Stepwise DIF Test

Step	Model	Description	LL	npar	Comparison	LRTS	df	p
2	M_2.1.11	#11: Non U DIF	-1471.455	10				
	M_2.0.12	#12: No DIF	-1701.334	7	M_2.0.12 vs M_2.1.12	15.950	3	0.001
	M_2.1.12	#12: Non U DIF	-1693.359	10				
	M_2.0.13	#13: No DIF	-1599.589	7	M_2.0.13 vs M_2.1.13	19.798	3	<0.001
	M_2.1.13	#13: Non U DIF	-1589.690	10				
	M_2.0.14	#14: No DIF	-1642.610	7	M_2.0.14 vs M_2.1.14	8.894	3	0.031
	M_2.1.14	#14: Non U DIF	-1638.163	10				
3	M_3.0	all Non U DIF Items			M_1.0 vs M_3.0			
					M_3.0 vs M_1.0			
4	M_4.1	#2 (U DIF) All other (Non U DIF)	-7746.505	65	M_4.1 vs M_3.0	0.366	2	0.416
	M_4.2	#3 (U DIF) All other (Non U DIF)	-7748.457	65	M_4.2 vs M_3.0	4.270	2	0.059
	M_4.3	#5 (U DIF) All other (Non U DIF)	-7747.987	65	M_4.3 vs M_3.0	3.330	2	0.094
	M_4.4	#11 (U DIF) All other (Non U DIF)	-7746.532	65	M_4.4 vs M_3.0	0.420	2	0.405
	M_4.5	#12 (U DIF) All other (Non U DIF)	-7747.913	65	M_4.5 vs M_3.0	3.182	2	0.102
	M_4.6	#13 (U DIF) All other (Non U DIF)	-7752.585	65	M_4.6 vs M_3.0	12.526	2	0.001
	M_4.7	#14 (U DIF) All other (Non U DIF)	-7750.864	65	M_4.7 vs M_3.0	9.084	2	0.005
5	M_5.0	#13, 14 (Non U DIF) #2, 3 ,5, 11, 12 (U DIF)	-7752.105	57	M_5.0 vs M_3.0	11.566	10	0.072
6	M_6.0	C on Country @ 0	-7787.161	55	M_6.0 vs M_6.1	270.124	2	<0.001
	M_6.1	C on Country (free)	-7752.105	57				

Step 3: To identify the optimal model, the fit of model M_3.0, where the seven identified items with DIF (2, 3, 5, 11, 12, 13, 14) were simultaneously set as non-uniform DIF, was compared with that of the baseline model M_1.0 (No_DIF) and model M_1.1 (All_DIF). The results revealed that the fit of model M_3.0 was significantly better than that of the baseline model M_1.0. Moreover, although the model where all items were set as DIF (M_1.1) exhibited better fit compared to M_3.0, the difference in fit between these two models was not substantial. Considering the improvement in fit from M_1.0 to M_3.0 and the parsimony of the model, model M_3.0 was chosen as the optimal latent class MIMIC model to proceed to the next step.

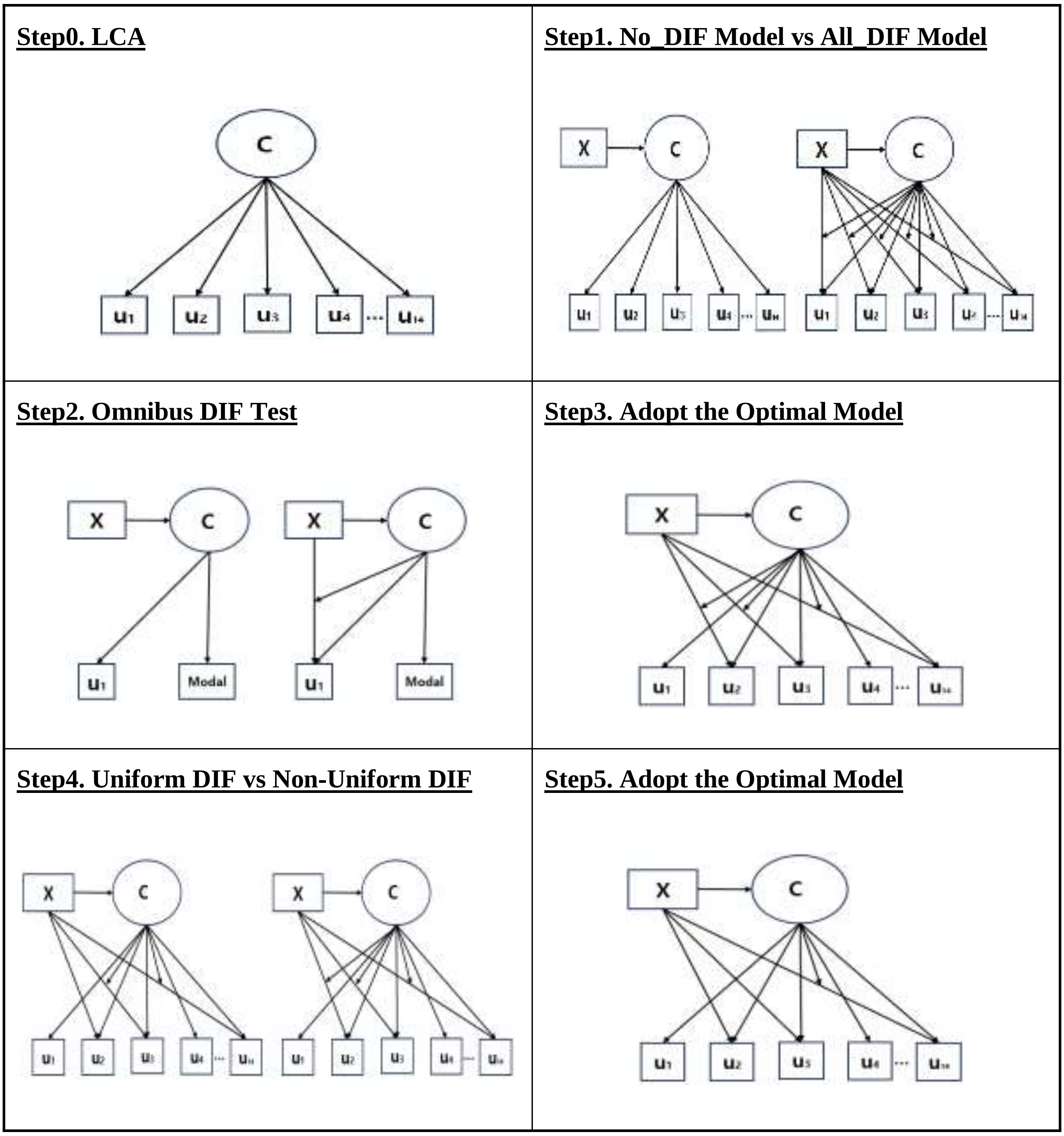
Step 4: To determine whether the seven identified DIF items were uniform DIF, the fit of model M_4.X, where the direct effects of the country variable were constrained to be uniform across classes for each of the seven items, was compared with that of model M_3.0, where all DIF items were treated as non-uniform DIF. As a result, items 2, 3, 5, 11, and 12 were confirmed to be uniform DIF, while items 13 and 14 were confirmed to be non-uniform DIF.

Step 5: The fit of model M_5.0, where the effects of the country on uniform DIF items were constrained to be uniform across classes, was compared with that of model M_3.0. As shown in Table 4, the fit of model M_5.0 was not significantly worse than that of M_3.0, indicating that the imposition of uniform DIF constraints did not significantly deteriorate the fit of the model. Therefore, model M_5.0 was adopted as the optimal model for the next step.

Step 6: Finally, model M_5.0 was re-designated as model M_6.1, and in model M_6.0, the polynomial logistic slope for the effect of the country variable on latent class membership was fixed to 0. In other words, while model M_6.1 allowed the country to freely estimate latent class membership, model M_6.0 did not allow the estimation of latent class membership by the country. Additionally, models M_6.0 and M_6.1 included all uniform and non-uniform DIF effects. The fit of models M_6.0 and M_6.1 was compared, and as shown in Table 4, the fit of model M_6.1, which allowed the country to freely estimate latent class membership, was significantly better than that of M_6.0, indicating the association of the country with latent class membership. Thus, the final adopted latent class MIMIC model, M_6.1, is illustrated in Figure 5, Step 5.

Figure 5

Latent Class MIMIC Modeling



Interpretation of the Final Model

An examination of the composition of Korean and Turkish nationals across the latent classes of mathematics achievement in the ultimately adopted M_6.1 model revealed that the high-achievement group comprised 81.6% Korean and 18.4% Turkish nationals, while the moderate-achievement group consisted of 59.3% Korean and 40.7% Turkish nationals. Furthermore, the low-achievement group comprised 21.2% Korean and 78.8% Turkish nationals.

The seven items of the mathematics achievement test identified as exhibiting DIF effects in the M_6.1 model, along with their respective effect sizes, are presented in Table 5 and Table 6. First, when examining the items identified as exhibiting uniform DIF, item 2 uniformly favored Korea across all classes, with a large DIF effect size. Item 3 similarly favored Korea uniformly across all classes, albeit with a negligible DIF effect size. Conversely, item 5 uniformly favored Türkiye across all classes, with a moderate DIF effect size, while item 11 favored Korea uniformly across all classes, with a large DIF effect size. Additionally, item 12 uniformly favored Türkiye across all classes, with a moderate DIF effect size.

Considering items identified as displaying non-uniform DIF, item 13 exhibited a significant in favored of Türkiye with a large effect size in the moderate-achievement group, while the DIF effects were not significant in the high- and low-achievement groups. On the other hand, item 14 favored Korea significantly with a large effect size in the low-achievement group, while the DIF effects were not significant in the high- and moderate-achievement groups.

Furthermore, Table 7 presents the results of logistic regression analysis on the influence of the country variable on the membership of latent classes of mathematics achievement. In Korea, there was a clear tendency for individuals to belong to either the high-achievement group or moderate-achievement group rather than the low-achievement group. Moreover, individuals in Korea were more likely to be part of the high-achievement group than the moderate-achievement group.

Table 5

Uniform DIF

Uniform DIF					
Item	**Est**	**SE**	**Est/SE**	**p**	**Effect size**
# 2	0.942	-0.220	4.274	<0.001	Large
# 3	0.427	0.176	2.422	0.015	Negligible
# 5	-0.526	0.172	-3.052	0.002	Moderate
# 11	1.980	-0.188	10.517	<0.001	Large
# 12	-0.455	0.177	-2.570	0.010	Moderate

Table 6

Non-Uniform DIF

Non-Uniform DIF									
	High group			**Moderate group**			**Low group**		
Item	**Est**	**p**	**Effect**	**Est**	**p**	**Effect**	**Est**	**p**	**Effect**
# 13	-1.049	0.310	Large	-1.123	0.011	Large	0.408	0.101	Negligible
# 14	0.435	0.406	Moderate	-0.010	0.967	Negligible	1.131	<0.001	Large

Table 7

Logistic regression analysis

	High group	Moderate group	Low group
Country Est (odds ratio)	2.838(17.084)	1.680(5.363)	Ref
	1.159(3.186)	Ref	

Conclusion and Discussion

Mathematics is a subject that significantly impacts students' academic success and future careers. Many countries participate in international academic achievement assessment to compare their performance with other nations and to explore the factors that influence academic achievement. Korea is the top performing country in the TIMSS 2019 math test, while Türkiye, although performing around the international average, has shown a steady increase in its performance since participating in TIMSS. Additionally, Türkiye has a more positive attitude towards mathematics compared to Korea. This study employed the LCA MIMIC method proposed by Masyn (2017) to explore the heterogeneous latent classes of mathematics achievement in the TIMSS 2019 assessment among 8th-grade students in Korea and Türkiye. Subsequently, the DIF of the mathematics assessment was examined according to country (Korea/Türkiye) to explore measurement invariance. The influence of the national variable (Korea/Türkiye) on membership in the latent classes of mathematics achievement was then investigated. The conclusions of this study are as follows:

First, a latent class analysis of mathematics achievement was conducted, identifying three distinct latent classes among the combined group of Korean and Turkish students: high-achievement, moderate-achievement, and low-achievement. The high-achievement group exhibited high item difficulty index of 0.7 or above for most items, with a higher proportion of Korean students in the group compared to Turkish students. The moderate-achievement group showed a wide range of item difficulty index varying from 0.03 to 0.95 across items, and exhibited a difficulty index below 0.1 in some geometry-related items, with a higher proportion of Korean students in the group compared to Turkish students. The low-achievement group demonstrated a consistently low item difficulty index of 0.4 or below for most items, with a higher proportion of Turkish students in the group compared to Korean students.

Secondly, in exploring DIF to verify the measurement invariance of mathematics achievement test items between Korea and Türkiye, a total of 7 out of 14 items were identified as exhibiting DIF. Among these, some items were identified as displaying uniform DIF, while others showed non-uniform DIF. This indicates the presence of direct effects of the country on individual items within the detected latent classes of mathematics achievement, and these direct effects were observed to vary in their application across latent classes, either uniformly or non-uniformly. Notably, while the overall item difficulty index for items indicated higher performance for Korea compared to Türkiye, this study's exploration of heterogeneous latent classes of mathematics achievement and subsequent examination of DIF based on country within these identified classes revealed areas of favorable or unfavorable performance in mathematics between Korea and Türkiye within homogeneous characteristics and ability groups. Furthermore, these results demonstrate that when analyzing the effects of covariates on latent classes, ensuring unbiased results requires conducting measurement invariance tests to confirm the direct effects of covariates on indicator variables.

Third, out of the seven items identified as DIFs, five items were identified as uniform DIFs and two items were identified as non-uniform DIFs. For items 2, 3, 5, 11, and 12, which exhibited uniform DIF, items 2, 3, and 11 favored Korean students in all classes, with large, negligible, and large DIF effect sizes, respectively. Additionally, items 5 and 12 favored Turkish students in all classes, with moderate DIF effect sizes for both items. Next, for items 13 and 14, identified as non-uniform DIF, item 13 favored Turkish students with a large effect size in the moderate-achievement group, while the DIF effect was not significant in the high- and low-achievement groups. Conversely, item 14 favored Korean students in the low-achievement group with a large effect size, with no significant effect observed in the high- and moderate-achievement groups. Additionally, examining the pattern of uniform/non-uniform DIF based on content areas, it was found that items in the Number, Geometry, and Algebra domains exhibited uniform DIF, whereas items in the Data and Probability domains displayed non-uniform DIF effects. Thus, it was observed that DIF effects varied between uniform and non-uniform across different content areas in mathematics.

Fourth, excluding items with non-significant or negligible DIF effects, examining the mathematics content domain and cognitive domains of items 2, 5, 11, 12, 13, and 14, which exhibit moderate or higher DIF effect sizes, it is found that items 2 and 11 correspond to the Number/Knowing and Geometry/Reasoning domains, respectively, and favor Korea across all latent classes. Item 14 corresponds to the Data and Probability/Applying domain and favors Korea in the low-achievement group. Items 5 and 12 correspond to the Algebra/Applying and Data and Probability/Applying domains, respectively, and favor Türkiye across all latent classes. Additionally, item 13 corresponds to the Data and Probability/Knowing domain and favors Türkiye in the moderate-achievement group. Summarizing the favorable and unfavorable items by country, Korea has one favorable item each in the Number/Knowing, Geometry/Reasoning, and Data and Probability/Applying domains, while Türkiye has one favorable item each in the Algebra/Applying, Data and Probability/Applying, and Data and Probability/Knowing domains. These results differ somewhat from Şen and Arıcan (2015), who reported that Korean students outperformed Turkish students in most math content domains (Number, Algebra, Geometry, Data and Probability). The reason for this partial discrepancy with Şen and Arıcan's (2015) study is that this study classified all students in Korea and Türkiye into heterogeneous latent classes based on their math achievement. It identified areas of favorability or unfavorability in math tests for homogeneous ability groups in Korea and Türkiye by exploring DIF within homogeneous latent classes. In particular, the results of this study showed that in the Data and Probability domain, Korea had a favorable result on one item compared to Türkiye in the low-achievement group. However, Türkiye had a favorable result on one item in each of the latent classes and in the moderate-achievement group compared to Korea. These findings align with Yoon and Lee's (2013) study, which reported that Korean students exhibited unfavorable performance in the Data and Probability domain compared to American students. This was evidenced by the exploration of DIF in the TIMSS 2007 assessment. Thus, it can be inferred that within homogeneous achievement groups, Korean students' performance in the Data and Probability domain is somewhat lower compared to that of Turkish students.

Fifth, examining the distribution of students across math achievement latent classes in each country, 41.8% of Korean students are classified as the high-achievement group, 40.5% as the moderate-achievement group, and 17.7% as the low-achievement group. In contrast, 9.1% of Turkish students are in the high-achievement group, 26.3% are in the middle-achievement group, and 64.6% are in the low-achievement group. The multinomial logistic regression analysis examined the impact of the country (Korea/Türkiye) on latent class membership in math achievement. The results indicated that students from Korea were more likely to be part of the high- and moderate-achievement groups rather than the low-achievement group. Additionally, students in Korea were more likely to be in the high-achievement group than in the moderate-achievement group.

In this study, the relationship between country (Korea/Türkiye) and membership in latent classes of math achievement was examined. To ensure the validity and robustness of the results, measurement invariance tests, including the detection of differential item functioning (DIF), were conducted. These tests were crucial for providing unbiased results in the identification of latent classes and assessing the impact of covariates in the LCA. Through the examination of measurement invariance for indicator variables and DIF in LCA, it was possible to identify areas of favorability or unfavorability across countries for individual items in mathematics achievement tests within homogeneous ability groups. Particularly noteworthy is the utilization of the MIMIC model in LCA for exploring DIF, which differs from previous studies (Kalaycioğlu & Berberoğlu, 2011; Lyons-Thomas et al., 2014; Yildirim, 2006) that applied classical test theory, item response theory, and logistic regression analysis in the exploration of DIF. Subsequent research can identify the causes of favorable or unfavorable areas in math achievement tests by country through a content-based approach to mathematics education. This can provide insights for enhancing the curriculum and educational methods within each country's mathematics education system.

Declarations

Author Contribution: Hyo Seob SONG: 1st Author, conceptualization, methodology, data analysis, writing & editing. Hee Sun JUNG: Corresponding Author, investigation, data analysis, visualization, supervision, writing - review & editing.

Ethical Approval: All ethical guidelines for authors have been followed. Ethical approval is not required for this study as it utilizes publicly available data.

Conflict of Interest: The authors declare no potential conflicts of interest.

Funding: No funds, grants, or other support were received during the preparation of this manuscript.

References

Arıkan, S., Van de Vijver, F., & Yagmur, K. (2016). Factors contributing to mathematics achievement differences of Turkish and Australian students in TIMSS 2007 and 2011. *EURASIA Journal of Mathematics, Science and Technology Education*, *12*(8), 2039-2059. https://doi.org/10.12973/eurasia.2016.1268a

Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using M plus. *Structural equation modeling: A multidisciplinary Journal*, *21*(3), 329-341. https://doi.org/10.1080/10705511.2014.915181

Badri, M. (2019). School Emphasis on Academic Success and TIMSS Science/Math Achievements. *International Journal of Research in Education and Science*, *5*(1), 176-189. https://eric.ed.gov/?id=EJ1197994

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

Demirus, K. B., & Pektas, S. (2022). Investigation of Timss 2015 science test items in terms of differential item functioning according to language and culture. *International Journal of Education Technology & Scientific Researches*, *7*(18), 1166-1178. http://dx.doi.org/10.35826/ijetsar.499

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

Dittrich, K., & Neuhaus, D. A. (2023). Korea's 'education fever' from the late nineteenth to the early twenty-first century. *History of Education*, *52*(4), 539-552. https://doi.org/10.1080/0046760X.2022.2098391

Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and standardization 1, 2. *ETS Research Report Series*, *1992*(1), i-40. https://doi.org/10.1002/j.2333-8504.1992.tb01440.x

Geesa, R. L., Izci, B., Song, H., & Chen, S. (2019). Exploring Factors of Home Resources and Attitudes Towards Mathematics in Mathematics Achievement in South Korea, Turkey, and the United States. *Eurasia Journal of Mathematics, Science and Technology Education*, *15*(9), eM_1751. https://doi.org/10.29333/ejmste/108487

Geesa, R. L., Izci, B., Chen, S., & Song, H. S. (2020). The Role of Gender and Attitudes toward Science in Fourth and Eighth Graders' Science Achievement in South Korea, Turkey, and the United States. *Journal of Research in Education*, *29*(2), 54-87. https://eric.ed.gov/?id=EJ1274027

Guhl, P. (2019). The impact of early math and numeracy skills on academic achievement in elementary school.

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European journal of psychological assessment*, *17*(3), 164-172. https://doi.org/10.1027/1015-5759.17.3.164

Huang, J. (2020). *Assessing robustness of the Rasch mixture model to detect differential item functioning: A Monte Carlo simulation study* (Doctoral dissertation, University of Denver). https://digitalcommons.du.edu/etd/1784/

Hwang, H. J., & Hwang, H. K. (2008). An Effect of the Constructivist Discussion on Learning Attitude in Mathematics and Children's Mathematics Achievement. *Education of Primary School Mathematics*, *11*(1), 59-74. https://koreascience.kr/article/JAKO200801440607130.page

Im, S., & Park, H. J. (2010). A comparison of US and Korean students' mathematics skills using a cognitive diagnostic testing method: linkage to instruction. *Educational Research and Evaluation*, *16*(3), 287-301. https://doi.org/10.1080/13803611.2010.523294

Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity, *Marketing Science 16*(1), 39-59. https://doi.org/10.1287/mksc.16.1.39

Kalaycioğlu, D. B., & Berberoğlu, G. (2011). Differential item functioning analysis of the science and mathematics items in the university entrance examinations in Turkey. *Journal of Psychoeducational Assessment*, *29*(5), 467-478. https://doi.org/10.1177/0734282910391623

Klieme, E., & Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *European journal of psychology of education*, *16*, 385-402. https://link.springer.com/article/10.1007/BF03173189

Ko, D. H., & Jung, H. S. (2020). Analysis on the Effect of Mathematics Class Characteristics and Mathematical Confidence on Mathematical Academic Achievement: Applying Hierarchical Linear Model. *School Mathematics*, *22*(2), 313-332. https://doi.org/10.29275/sm.2020.06.22.2.313

Lee, J., & Stankov, L. (2018). Non-cognitive predictors of academic achievement: Evidence from TIMSS and PISA. *Learning and Individual Differences*, *65*, 50-64. https://doi.org/10.1016/j.lindif.2018.05.009

Lubinski, D., Benbow, C. P., & Kell, H. J. (2014). Life paths and accomplishments of mathematically precocious males and females four decades later. *Psychological Science*, *25*(12), 2217-2232. https://doi.org/10.1177/0956797614551371

Lyons-Thomas, J., Sandilands, D., & Ercikan, K. (2014). Gender Differential Item Functioning in Mathematics in Four International Jurisdictions. *Education & Science/Egitim ve Bilim*, *39*, 20-32.

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. https://doi.org/10.1080/10705511.2016.1254049

Mullis, I. V., & Martin, M. O. (2017). *TIMSS 2019 Assessment Frameworks*. International Association for the Evaluation of Educational Achievement. https://eric.ed.gov/?id=ed596167

Mullis, I. V., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 international results in mathematics and science*. https://www.skolporten.se/app/uploads/2020/12/timss-2019-highlights-1.pdf

Nylund-Gibson, K., & Choi, A. Y. (2018). Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science*, *4*(4), 440-461. https://doi.org/10.1037/tps0000176

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: A Multidisciplinary Journal*, *23*(6), 782-797. https://doi.org/10.1080/10705511.2016.1221313

Ram, N., & Grimm, K. J. (2009). Methods and measures: Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. *International journal of behavioral development*, *33*(6), 565-576. https://doi.org/10.1177/0165025409343765

Saatcioglu, 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

Samuelsen, K. M. (2008). Examining differential item functioning from a latent mixture perspective. *Advances in latent variable mixture models*, 177-197.

Şen, S., & Arıkan, M. (2015). A diagnostic comparison of Turkish and Korean students' mathematics performances on the TIMSS 2011 assessment. *Journal of Measurement and Evaluation in Education and Psychology*, *6*(2). https://doi.org/10.21031/epod.65266

Shin, K., Jahng, K. E., & Kim, D. (2019). Stories of South Korean mothers' education fever for their children's education. *Asia Pacific Journal of Education*, *39*(3), 338-356. https://doi.org/10.1080/02188791.2019.1607720

Sohn, W. S. (2010). Exploring Potential Sources of DIF for PISA 2006 Mathematics Literacy Items: Application of Logistic Regression Analysis. *Journal of Educational Evaluation*, *23*(2), 371-390. https://scholar-kyobobook-co-kr-ssl.ca.skku.edu/article/detail/4010023071385

Song, H. S., Kim, H. C., & Jung, H. S. (2023). The Effect of Participation in Private Education on Math Affective Attitudes: Measurement Invariance and Differential Item Functioning in Latent Class MIMIC Model. *Journal of Educational Evaluation*, *36*(4), 687-709. http://dx.doi.org/10.31158/JEEV.2023.36.4.687

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

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political analysis*, *18*(4), 450-469. https://doi.org/10.1093/pan/mpq025

Wang, X. S., Perry, L. B., Malpique, A., & Ide, T. (2023). Factors predicting mathematics achievement in PISA: a systematic review. *Large-scale Assessments in Education*, *11*(1), 24. https://doi.org/10.1186/s40536-023-00174-8

Wiberg, M. (2019). The relationship between TIMSS mathematics achievements, grades, and national test scores. *Education Inquiry*, *10*(4), 328-343. https://doi.org/10.1080/20004508.2019.1579626

Woo, H., & Hodges, N. N. (2015). Education fever: Exploring private education consumption motivations among Korean parents of preschool children. *Family and Consumer Sciences Research Journal*, *44*(2), 127-142. https://doi.org/10.1111/fcsr.12131

Yildirim, H. H. (2006). The DIF analysis of mathematics items in the international assessment programs. https://www.proquest.com/openview/37bc178a79e82e2c13ca84983403a39a/1?cbl=2026366&diss=y&pq-origsite=gscholar

Yoon, J. Y., & Lee, Y. S. (2013). A Study of DIF Analyses using TIMSS (2007) Mathematics Test Across South Korea, U. S., and Singapore. *Journal of Educational Evaluation*, *26*(2), 415-439. https://scholar-kyobobook-co-kr-ssl.ca.skku.edu/article/detail/4010023595562