Journal of Pedagogical Research
Onlinefirst Article, 2024
https://doi.org/10.33902/JPR.202426420

Research Article

Assessing the fairness of mathematical literacy test in Indonesia: Evidence from gender-based differential item function analysis

Kartianom Kartianom[1], Heri Retnawati[2] and Kana Hidayati[3]

[1] *Yogyakarta State University & IAIN Bone, Indonesia (ORCID: 0000-0002-4734-042X)*
[2] *Yogyakarta State University, Indonesia (ORCID: 0000-0002-1792-5873)*
[3] *Yogyakarta State University, Indonesia (ORCID: 0000-0002-9226-8500)*

Conducting a fair test is important for educational research. Unfair assessments can lead to gender disparities in academic achievement, ultimately resulting in disparities in opportunities, wages, and career choice. Differential Item Function [DIF] analysis is presented to provide evidence of whether the test is truly fair, where it does not harm or benefit certain groups of students. For this reason, this study aims to assess the fairness of mathematics literacy tests from a gender perspective using three DIF analysis approaches, namely, the Cognitive Diagnostic Model [CDM], Classical Test Theory [CTT], and Item Response Theory [IRT], and to compare the results of the three approaches to examine the compatibility between them in identifying DIF effects. This study was included in quantitative descriptive research, and for the CDM approach, a retrofitting method (post-hoc analysis) was used. The sample in this study consists of Indonesian students who participated in the administration of PISA 2012 and were tested on Booklet 1, Booklet 3, Booklet 4, and Booklet 6. The Q-matrix used in this study consisted of 12 items and 11 attributes. The results of this study show that out of the 12 items analyzed, there are differences in findings between the CTT, IRT, and CDM approaches; the item with the largest DIF was found using the Raju Unsigned Area Measures method in IRT and the Wald Test from the CDM approach, while the item with the lowest DIF was found using the LRT method from the CDM approach; and there are three items that were simultaneously identified as DIF using the CTT, IRT, and CDM methods, namely PM923Q01, PM923Q03, and PM924Q02. Items PM923Q01 and PM923Q03 favor the group of male students, while item PM924Q02 favors the group of female students.

Keywords: Classical test theory; Cognitive diagnostic model; Differential item function; Fairness; Item response theory; PISA 2012

Article History: Submitted 5 February 2024; Revised 7 May 2024; Published online 26 June 2024

1. Introduction

The Program for International Student Assessment [PISA] is a comprehensive educational assessment program initiated by the Organization for Economic Cooperation and Development [OECD] in 1997. PISA is conducted every three years to evaluate the essential knowledge and critical skills necessary for approximately 15-year-old students to succeed in society (OECD, 2018). PISA has had a significant impact on educational practices and reforms in numerous countries, particularly in the formation of national policies (Wu et al., 2020). In recent years, researchers have

Address of Corresponding Author

Kartianom Kartinom, Yogyakarta State University, Daerah Istimewa Yogyakarta 55281, Indonesia.

✉ kartianom.2022@student.uny.ac.id

How to cite: Kartianom, K., Retnawati, H., & Hidayati, K. (2024). Assessing the fairness of mathematical literacy test in Indonesia: Evidence from gender-based differential item function analysis. *Journal of Pedagogical Research*. Advance online publication. https://doi.org/10.33902/JPR.202426420

advocated new methods of measurement to present and comprehend PISA outcomes (Rutkowski & Rutkowski, 2016). The Cognitive Diagnostic Model (CDM) is considered suitable for these needs, as it combines modern statistical methods with cognitive theory (Wu et al., 2020). CDM is a psychometric model that provides specific information about the interrelated yet separable mastery of attributes by test-takers (Hou et al., 2014). Compared to Item Response Theory (IRT), which positions students on a continuous latent variable, categorical latent attributes are predicted by CDM, which refers to the skills and abilities that underlie the construction of items (Paulsen et al., 2020; Ravand & Baghaei, 2020).

In addition to providing more detailed information about test participants, CDM is capable of classifying test participants according to their mastery profiles (Rupp et al., 2010). Correct test participant responses indicate attribute mastery, represented by "1" in the Q-Matrix, while incorrect responses are represented by "0" (Ravand & Baghaei, 2020; Rupp et al., 2010). In CDM, the Q-Matrix serves as the key to determining the mastery profile of test participants, as it contains the attributes planned to be measured by the test (Eren et al., 2023; Li & Traynor, 2022). The Q-Matrix represents the underlying attributes in a multidimensional format and organizes them into rows and columns, with attributes placed in the columns and underlying attributes in the rows (Li & Traynor, 2022; Rupp et al., 2010). The alignment of items and attributes in the matrix enhances the validity of interpreting the scores collected from student responses (Li & Traynor, 2022; Ravand & Baghaei, 2020). The process of developing the Q-matrix used in CDM is fundamental for the development of diagnostic tests (Kang et al., 2019). When this step is neglected, there is potential for bias in the test and issues in student classification (De La Torre & Chiu, 2016). Bias in tests can lead to serious effects for both individuals and society (Moradi et al., 2016). Therefore, sources of bias must be identified and addressed in the test design and score interpretation to ensure fairness in testing.

The implementation of a fair test is important in educational research. Moradi et al. (2016) demonstrate that unfair assessment can lead to gender disparities in academic achievement, ultimately resulting in unequal opportunities, wages, and careers. Therefore, it is necessary to conduct a differential item function analysis using the CDM approach to ensure test fairness. This procedure can be valuable for identifying items that contain bias (Hou et al., 2014; Ma et al., 2021; Paulsen et al., 2020). Typically, DIF is understood to be a consequence of disparities in the likelihood of answering a question accurately among students who possess the same level of proficiency but originate from diverse cohorts (Hou et al., 2014; Mehrazmay et al., 2021). For example, a particular item may disproportionately benefit a male group in terms of its construct. This item may have been influenced by bias due to a disparity in the item response function between the male and female groups. If not addressed, this could affect the accuracy of the score interpretation derived from the test, as other constructs could become mixed with the intended constructs being assessed (Eren et al., 2023). Therefore, it is important to identify and examine items that contain bias to avoid issues and carry out appropriate measurement procedures.

Currently, Differential Item Functioning (DIF) is a standard procedure that must be conducted in psychometric analysis. The DIF in the context of CDM holds the same significance as in the traditional approach. The conventional method evaluates a person's inherent ability, whereas CDM assesses the change in the likelihood of correctly answering a question for individuals with the same attribute profile, regardless of the group they belong to (Wu et al., 2020). The meaning of DIF, as seen from the CDM perspective, is the impact of varying the likelihood of correctly answering a question among students from diverse groups with the same proficiency level (Hou et al., 2014). The presence of items containing DIF can undermine prediction accuracy and disrupt attribute profiles (Paulsen et al., 2020). The effect of DIF is harmful when it comes to contrasting hidden classes among various groups (Eren et al., 2023). Furthermore, DIF analysis plays a crucial role in testing parameter invariance (Ma et al., 2021). According to the attribute profiles, invariance occurs when item responses are conditioned independently. Therefore, DIF analysis becomes an important factor in determining whether the attribute-item interaction between groups is invariant.

Previous research focusing on the determination of DIF in CDM is still very limited. The Wald test method, developed by Hou et al. (2014), was designed to detect both uniform and non-uniform DIF within the CDM framework. Liu et al. (2019) applied covariance matrices to examine the performance of the Wald test method in identifying DIF. One year later, Hou et al. (2020) employed the Wald test equation to determine DIF in the CDM. Additionally, Akbay (2021) utilized three methods for assessing DIF, namely, the Classical Test Theory (MH) approach, Item Response Theory (Raju), and CDM (Wald Test), to investigate the psychometric attributes of the test. Large-scale assessment data were used to observe DIF determination patterns using the three DIF detection methods. The data were collected using two types of booklets. DIF analysis was conducted based on the variables of class type and booklet type. Unlike Ma et al. (2021), who adapted the premises of the G-DINA model to create the multi-group G-DINA (MG G-DINA) model for detecting DIF, the MG G-DINA model can distinguish among diverse student groups and their varying utilization of shared or distinct attributes in various manners. Moreover, a comparison of the model's performance was conducted using the likelihood ratio test (LRT) and Wald test.

Although methods for determining Differential Item Functioning (DIF) are already available, existing literature indicates that a more effective approach to estimating DIF still needs to be investigated. Most research related to CDM comes from non-diagnostic tests, which raise many psychometric questions about DIF in CDM research, but they have yet to be answered. Non-diagnostic research integrated into retrofitting (CDM) aims to provide detailed information on students' strengths and weaknesses. This is an important step in the transition from single-score reporting to CDM, which provides holistic feedback (Wu et al., 2020). The retrofitting approach is beneficial for determining DIF because it can provide detailed information about students' mastery based on the underlying attributes of item scoring (Terzi & Sen, 2019). Additionally, the retrofitting approach provides evidence for the validity and reliability of score interpretation, ensuring proper usage and interpretation in the CDM (Eren et al., 2023). Seeking meaning in evaluation without considering validity aspects will not provide benefits (Ma et al., 2021). When conducting CDM analysis with a large dataset, it is crucial to consider the validity of the analysis, as language, cultural, or demographic factors, such as gender, can impact student performance and lead to variations in the results.

Based on the considerations regarding the contributions of DIF determination methods and validity aspects described in previous research, CDM has significant potential and can provide confidence for it, provided that the methodology is correct. In this study, six DIF techniques/methods were used to assess the fairness of the mathematics literacy test on a large-scale assessment dataset from the PISA 2012. These techniques include the Wald test (Hou et al., 2020; Ma et al., 2021) and LRT (Ma et al., 2021) for the CDM approach; the Mantel-Haenszel method (MH) (Mantel & Haenszel, 1959) and logistic regression (LR) (Swaminathan & Rogers, 1990) for the CTT approach; and Lord's x^2 method (Lord, 1980) and Raju's unsigned area measures (Raju, 1988) for the IRT approach. The DIF analysis results from these six methods were then compared to examine the compatibility between the approaches for identifying DIF effects. For this purpose, previous research has shown that items containing DIF are frequently found in gender group bias analyses, particularly in numerical domains, such as mathematics (Başman & Kutlu, 2020; Eren et al., 2023; Ong et al., 2015; Wu et al., 2020; Yildirim, 2019). Therefore, in this study, gender is used as the DIF variable.

2. Method

2.1. Research Design

This study is quantitative descriptive research because the purpose of this study is to assess the fairness of mathematical literacy tests from a gender perspective using three DIF analysis approaches, CDM, CTT, and IRT, and to compare the results of the three approaches to examine the compatibility between the approaches in identifying DIF effects. For the CDM approach, a retrofitting method (post-hoc analysis) is employed, which involves extracting response data from non-diagnostic test instruments to obtain richer information. Additionally, the retrofitting approach is used to transform non-diagnostic test instruments into diagnostic tests (Ravand & Baghaei, 2020).

2.2. Sample

The sample consisted of Indonesian students who participated in the PISA administration conducted by the OECD in 2012. There were 5,622 Indonesian students, comprising 2,860 female and 2,762 male students, who participated in PISA 2012. However, for this study, only Indonesian students tested on Booklet 1, Booklet 3, Booklet 4, and Booklet 6 were selected. Therefore, 1,696 Indonesian students, comprising 840 females and 856 males, were used in this study.

2.3. Cognitive Attributes and Q-Matrix Structure

Currently, a vast array of test items has been incorporated into PISA assessments. However, for mathematics, only test items published in 2012 are accessible, whereas no items are available for other years. In this study, the Q-Matrix developed by Wu et al. (2020) is utilized. The developed Q-Matrix consists of 12 test items with 11 attributes and has undergone rigorous validation (Wu et al., 2020). The dimensions of the attributes and the arrangement of the Q-Matrix are presented in Table 1 and Table 2.

Table 1
Cognitive Attribute Dimensions PISA 2012

Dimension and code	*Attribute*	*Definition*
Content		
N1	Change and relationships	The relationship between quantities can be depicted through the use of algebraic expressions, equations, inequalities, and functions, as well as graphical representations.
N2	Space and shape	The relationships between planes, points, lines, and surfaces in space are involved, as well as other related elements.
N3	Quantity	Enhancing the measurement of object attributes, relationships, situations, and entities by incorporating them into the world, and assessing, interpreting, and illustrating the various forms of measurement.
N4	Uncertainly and data	Perceiving change, probability and chance, representation, evaluation, interpretation of data related to uncertainty.
Process		
P1	Mathematization	Employing mathematical language to illustrate and clarify real-world issues while converting relevant data into mathematical measurements.
P2	Mathematical operation	Mathematical concepts, facts, procedures, and reasoning are used to identify, calculate, analyze, and solve problems.
P3	Mathematical reality	The capacity to utilize mathematical solutions to address practical issues and assess and draw conclusions from the outcomes.
Context		
C1	Personal	Involves personal scenarios, primarily focusing on individual activities, family, or peer interactions.
C2	Occupational	Involves personal scenarios, primarily focusing on individual activities, family, or peer interactions.
C3	Societal	Social issues are centered in the individual's society, with a focus on problems that are of a societal perspective.
C4	Scientific	Involves issues in the scientific category as well as topics related to science and technology.

In Table 1, the four attributes in the content dimension encompass nearly all mathematical content in the compulsory learning stage. The three attributes in the process dimension are aligned with those described by renowned mathematician Freudenthal (1972). Subsequently, the attributes in the context dimension encompass all fields that students might encounter in the future, and they are vital in preparing students to perceive the world through a mathematical lens (Wu et al., 2020).

Table 2
Q-Matrix

Item	*N1*	*N2*	*N3*	*N4*	*P1*	*P2*	*P3*	*C1*	*C2*	*C3*	*C4*
PM00FQ01	0	0	1	0	0	0	1	1	0	0	0
PM903Q03	1	0	0	0	0	1	0	0	1	0	0
PM918Q01	0	0	0	1	0	0	1	0	0	1	0
PM918Q02	0	0	0	1	0	0	1	0	0	1	0
PM918Q05	0	0	0	1	0	1	0	0	0	1	0
PM923Q01	0	0	1	0	0	1	0	0	0	0	1
PM923Q03	0	1	0	0	0	1	0	0	0	0	1
PM923Q04	1	0	0	0	1	0	0	0	0	0	1
PM924Q02	0	0	1	0	1	0	0	1	0	0	0
PM995Q01	0	1	0	0	0	1	0	0	0	0	1
PM995Q02	0	1	0	0	1	0	0	0	0	0	1
PM995Q03	0	0	1	0	1	0	0	0	0	0	1

In Table 2, the formed Q-Matrix structure has a Cronbach's alpha reliability of .61 and a marginal reliability of .60, when treated as a general test based on the CTT and IRT approaches. When treated as a diagnostic test, the reliability of the attributes was calculated based on the CDM approach. The reliability indices for each attribute are .63, .77, .89, .91, .86, .88, .85, .80, .95, .82, and .83.

2.4. Data Analysis

The initial step in conducting DIF analysis using the CTT, IRT, and CDM approaches is to determine the grouping variable, in this case, the gender variable, where female students are considered the reference group (R), and male students are the focal group (F). Subsequently, model fit was assessed using the IRT approach based on relative fit indices, item fit, and convergence issues. The IRT models utilized in this study included the Rasch, 1PL, 2PL, 3PL, and 4PL models. The selection of the best IRT model was based on the smallest relative fit index, highest number of well-fitting items, and absence of convergence problems. The selected IRT model was then used to test the assumptions of unidimensionality, local independence, and parameter invariance before proceeding to DIF analysis using the IRT approach. The same procedure was applied to the CDM approach before conducting the DIF analysis. In this study, CDM models focused on G-DINA and DINA. From these two models, the model fit was further assessed based on the relative fit indices and the assumption of local independence. The selection of the best CDM model is based on the smallest -2LL, AIC, BIC, and SABIC values and the Likelihood Ratio test (LR). The best model was then used to test the assumption of local independence based on the Max(X2) value. The assumption of local independence was satisfied if the p-value of Max(X2) was greater than .05. All analyses for the CTT, IRT, and CDM approaches are conducted using the R software with the packages "mirt" (Chalmers, 2012), "CDM" (George et al., 2016), "difR" (Magis et al., 2010), and "GDINA" (Ma & De La Torre, 2020). The Holm method, as suggested by Ma et al. (2021), was employed to correct the p-values from various methods to evaluate the DIF and control the Type-1 errors with an alpha of .05.

3. Results

3.1. Model IRT Fit

Based on the information presented in Table 3, the 2PL and 4PL models had the smallest relative fit indices compared to the other two IRT models. In the 2PL model, the smallest relative fit indices were observed for BIC and SABIC, whereas in the 4PL model, they were observed for -2LL and AIC. In terms of the number of well-fitting items, the 4PL model had the highest number compared to the other three IRT models. However, convergence issues were encountered when analyzing the data using the 3PL and 4PL models, which required a modification of the default TOL criterion from 0.0001 to 0.001. Considering convergence issues as part of the initial data and model fit detection in the IRT analysis, the 3PL and 4PL models were not used in this study. Therefore, the 2PL model was the best choice for conducting the DIF analysis in this study.

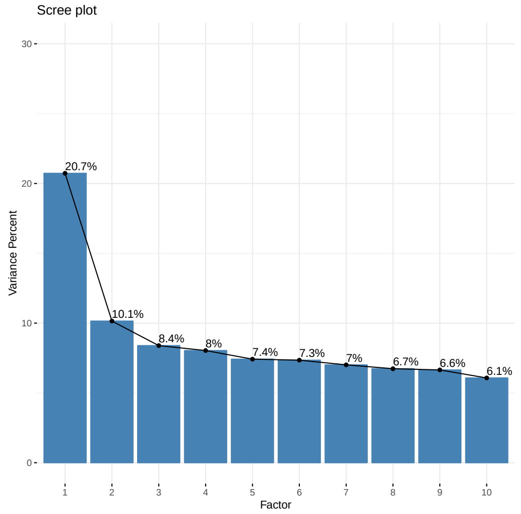
Table 3
Comparison of IRT Model Fit

Model	*-2LL*	*AIC*	*BIC*	*SABIC*	*Item Fit*	*Convergence*
Rasch/1PL	16774.42	16800.42	16871.09	16829.79	5	0.0001
2PL	16608.58	16656.58	16787.04	16710.80	10	0.0001
3PL	16560.08	16632.08	16827.78	16713.41	10	0.001
4PL	16526.59	16622.59	16883.52	16731.03	11	0.001

3.1.1. Assumptions of unidimensionality and local independence

Before testing the assumption of unidimensionality, the adequacy of the sample was assessed using Bartlett’s test. Based on the Bartlett test results, a KMO value greater than 0.5 was obtained, specifically 0.8, was obtained. This indicates that the sample used in this study was sufficient for factor analysis. The results of the factor analysis showed that the 12 PISA 2012 items used in this study measured a single dimension, namely, students' mathematical literacy. This can be observed in the Scree Plot shown in Figure 1, where the domain factor can explain 21% of the variance in the data. As the assumption of unidimensionality was fulfilled, the assumption of local independence was also automatically fulfilled.

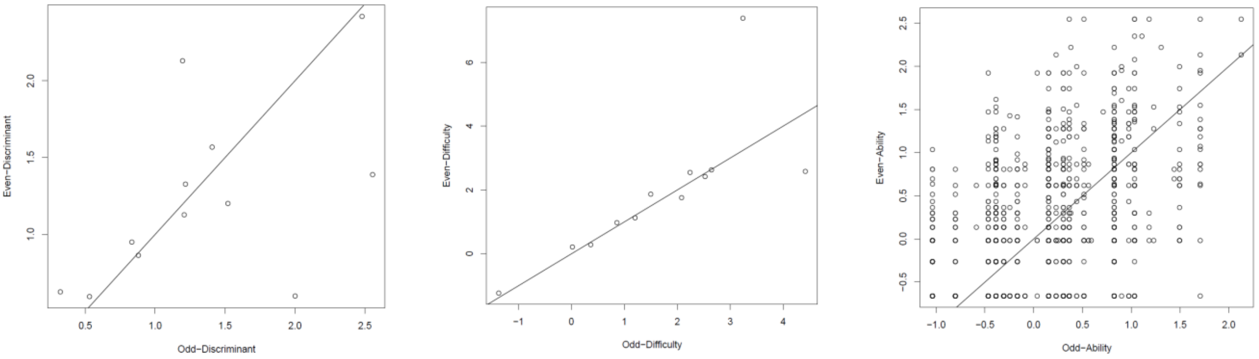
Figure 1
Scree Plot



3.1.2. Parameter invariance

The test of parameter invariance was based on the best IRT model obtained from the model fit test, namely, the 2PL model. Based on the information presented in Figure 2, both the item and ability parameters fulfilled the assumption of invariance. This can be observed in Figures 2.A and 2.B, where the points of item discrimination and item difficulty from odd- and even-numbered student groups follow a linear relationship (having a strong correlation). The same applies to Figure 2.C, where the points of student abilities when answering odd-numbered and even-numbered items also follow a linear relationship. Therefore, it can be concluded that the data in this study meet the assumption of parameter invariance based on the 2PL model.

Figure 2
Parameter Invariance



A. Discriminant Invariance **B. Difficulty Invariance** **C. Ability Invariance**

3.2. Model Fit of CDM

The selection of the CDM model was based on relative fit indices. According to the information presented in Table 4, the G-DINA model had the smallest values for the relative fit indices -2LL and AIC, whereas the DINA model had the smallest values for the BIC, CAIC, and SABIC indices. In this situation, a Likelihood Ratio (LR) test was conducted. The null hypothesis tested in the LR test is "H0: The fit of the reduced model (DINA) is as good as that of the more complex or saturated model (G-DINA).” The results of the LR test are presented in Table 4, as indicated by the significant *p*-values. Therefore, it can be concluded that the data are a better fit for the G-DINA model.

Table 4
Comparison of Model Fit with Relative Indices

Model	*-2LL*	*Deviance*	*AIC*	*BIC*	*CAIC*	*SABIC*	χ^2	*df*	*p-value*
G-DINA	16135.8	16135.8	20421.8	32050.9	34193.9	25242.9			
DINA	16516.4	16516.4	20658.4	31896.7	33967.7	25317.5	380.5	72	<.001
DINO	16485.6	16485.6	20627.6	31865.9	33936.9	25286.7	349.7	72	<.001
A-CDM	16259.6	16259.6	20449.6	31818.2	33913.2	25162.7	123.7	48	<.001

3.2.1. Assumption of CDM Model

The test of the local independence assumption in the CDM model is based on the fit of the data to the CDM model, namely, the G-DINA model. Based on the information presented in Table 5, the *p*-value of Max(X2) was greater than .05. This can be interpreted as the absence of local independence. In other words, the assumption of local independence in the G-DINA model was neither violated nor fulfilled.

Table 5
Local Independence of CDM Model

Type	*Value*	*p-value*
Max(X2)	4.715	1
Abs(fcor)	0.067	.196

3.3. DIF Method with CTT Approach

The DIF detection method with the CTT approach was performed using MH and LR techniques. In Table 6 and Figure 3, the DIF information is presented based on the findings using the MH technique. The information in Table 6 indicates that there are six items (PM918Q01, PM918Q02, PM918Q05, PM923Q01, PM923Q03, and PM924Q02) with *p*-values less than .05, indicating DIF. The information in Table 6 supports the findings in Figure 3, which also shows the same items containing DIF, where item 3 (PM918Q01) had the largest effect size, and item 5 (PM923Q03) had the smallest effect size.

In addition to providing information on the significance of DIF, Table 6 provides information regarding the magnitude and size of the DIF. There is one item that has a large effect size on DIF (level C), namely item PM923Q04, but it is not significant. Furthermore, there are two items with a moderate effect size on DIF (level B), namely PM918Q01 and PM995Q02, while the other four items fall under a category that can be ignored (level A). To determine which group is beneficial, it is necessary to consider the value of ΔMH. If ΔMH was positive (+), the focal group (women) was supported, whereas if it was negative (−), the reference group (men) was supported. Considering the value of ΔMH for items PM918Q01 and PM995Q02 would benefit female students.

Figure 3
DIF Results Using the MH Method

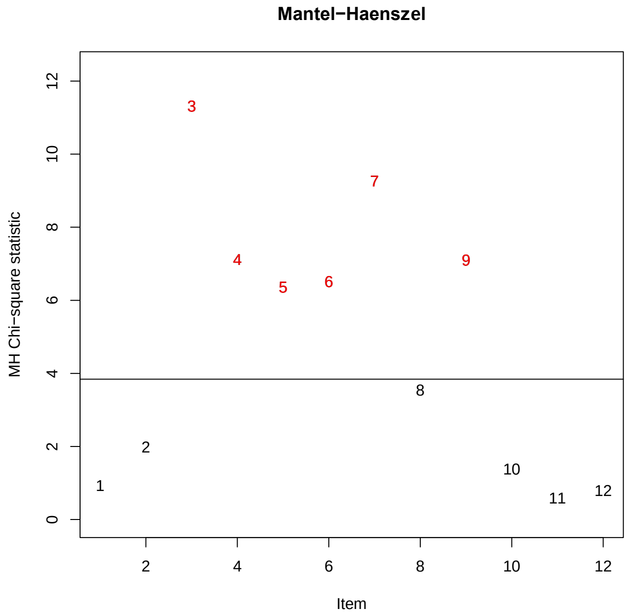


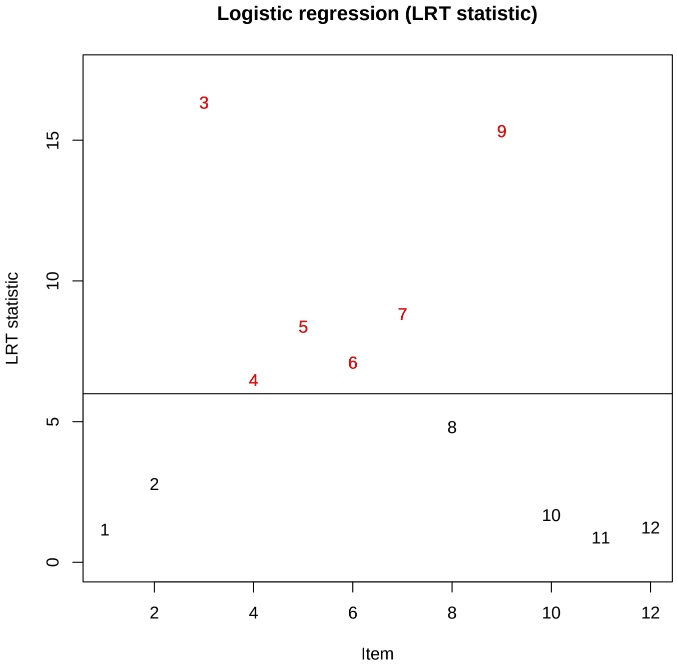
Table 6
DIF Results Using the MH Method

Item	χ^2	*p-value*	*Alpha MH*	*ΔMH*	*Effect Size*
PM00FQ01	0.9399	.3323	0.7999	0.5247	A
PM903Q03	1.9822	.1592	0.7581	0.6508	A
PM918Q01	11.3092	.0008	0.6306	1.0836	B
PM918Q02	7.1241	.0076	1.3973	−0.7862	A
PM918Q05	6.3601	.0117	0.7416	0.7025	A
PM923Q01	6.5056	.0108	1.4132	−0.8127	A
PM923Q03	9.2673	.0023	1.5106	−0.9694	A
PM923Q04	3.5471	.0596	2.5813	−2.2285	C
PM924Q02	7.0993	.0077	0.7034	0.8267	A
PM995Q01	1.3772	.2406	1.1933	−0.4152	A
PM995Q02	0.5889	.4428	0.6297	1.0869	B
PM995Q03	0.7935	.3730	1.2057	−0.4396	A

Note. Effect Size: 0 = A; 1.0 = B; 1.5 = C ‘A’: Negligible Effect; ‘B’: Moderate Effect; C = Large Effect.

Figure 4 and Table 7 present the results obtained using the LR technique. Based on the information obtained in Figure 4, six items (PM918Q01, PM918Q02, PM918Q05, PM923Q01, PM923Q03, and PM924Q02) were identified to contain DIF. Item 3 appears to deviate from the critical value, indicating the largest effect size of DIF, whereas Item 4 has the smallest effect size of DIF because the distance from the critical value is not significant.

Figure 4
DIF Results Using the LRT Method



When Table 7 was examined, six items containing DIF were found. The findings in Table 7 support those shown in Figure 4. However, based on the information in Table 7, these six items have DIF effect sizes that can be disregarded as they fall under level A.

Table 7
DIF Results Using the LRT Method

Item	χ^2	*p-value*	ΔR^2	*Effect Size*
PM00FQ01	1.1724	.5564	0.0015	A
PM903Q03	2.7952	.2472	0.0026	A
PM918Q01	16.3390	.0003	0.0074	A
PM918Q02	6.4806	.0392	0.0026	A
PM918Q05	8.3784	.0152	0.0037	A
PM923Q01	7.0921	.0288	0.0037	A
PM923Q03	8.8282	.0121	0.0047	A
PM923Q04	4.8066	.0904	0.0126	A
PM924Q02	15.3293	.0005	0.0066	A
PM995Q01	1.6702	.4338	0.0007	A
PM995Q02	0.8816	.6435	0.0042	A
PM995Q03	1.2326	.5399	0.0012	A

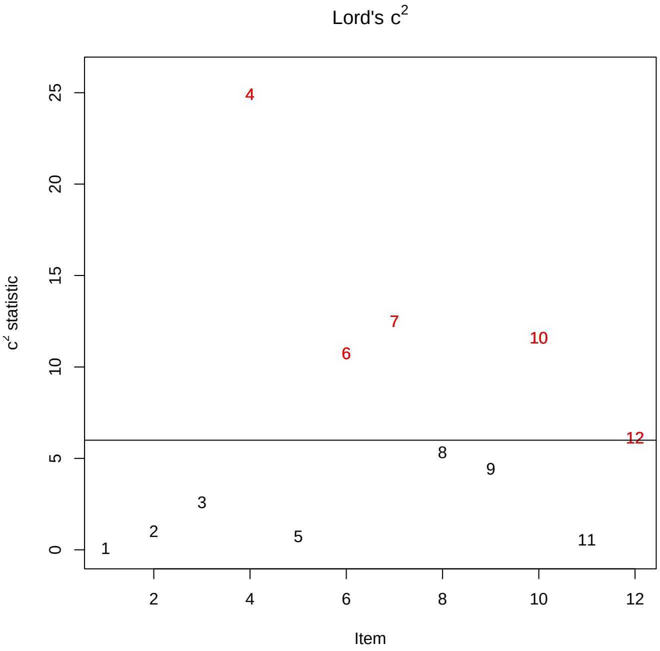
Note. Effect Size: 0.01 = A; 0.13 = B; 0.26 = C 'A': Negligible Effect; 'B': Moderate Effect; C = Large Effect.

3.4. DIF Method with IRT Approach

The detection of DIF using the IRT approach was carried out using Lord's χ^2 and Raju's unsigned area measurement techniques. The results of DIF detection using Lord's χ^2technique are presented in Figure 5 and Table 8. Figure 5 shows that five items colored red (4 = PM918Q02, 6 = PM923Q01, 7 = PM923Q03, 10 = PM995Q01, and 12 = PM995Q03) had values above the threshold, indicating the presence of DIF. Among these five items, it is evident that Item 4 (PM918Q02) is the farthest from the critical value, whereas Item 12 (PM995Q03) is not as far. This can be interpreted as the largest effect size of DIF being present in item PM918Q02, whereas the smallest effect size of DIF is found in PM995Q03.

Figure 5

DIF Results Using the Lord's χ^2 Method



When Table 8 is examined, it can be observed that the five items containing DIF in Figure 5 have *p*-values less than .05. Among these five items, two (4 = PM918Q02, 10 = PM995Q01) are also identified as having DIF based on Raju's Unsigned Area Measures technique presented in Figure 6 and Table 9.

Table 8

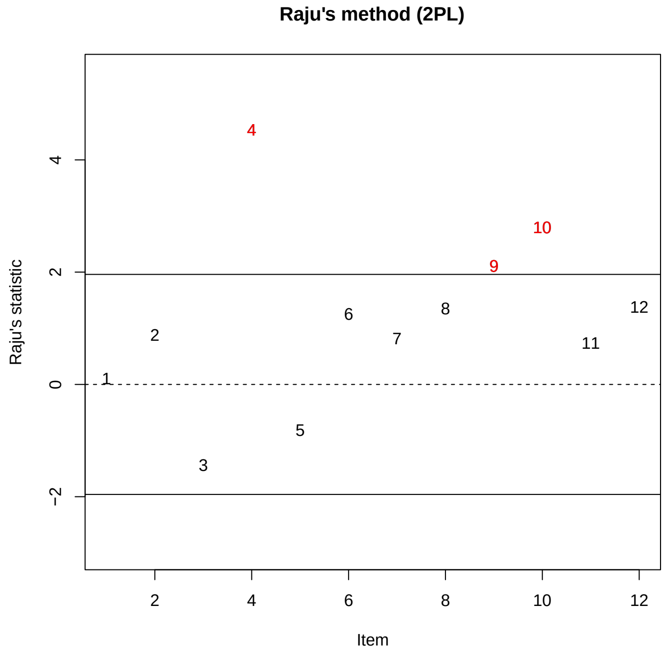
DIF Results Using the Lord's χ^2 Method

Item	χ^2	*p-value*
PM00FQ01	0.1028	.9499
PM903Q03	1.0221	.5999
PM918Q01	2.6000	.2725
PM918Q02	24.9053	.0000
PM918Q05	0.7413	.6903
PM923Q01	10.7500	.0046
PM923Q03	12.5041	.0019
PM923Q04	5.3428	.0692
PM924Q02	4.4387	.1087
PM995Q01	11.6050	.0030
PM995Q02	0.5593	.7561
PM995Q03	6.1220	.0468

Based on the information presented in Figure 6, three items colored in red (4 = PM918Q02, 9 = PM924Q02, and 10 = PM995Q01) were above the threshold, indicating the presence of DIF. Among these three items, it can be observed that Item 4 (PM918Q02) has the farthest distance from the critical value, whereas Item 9 (PM924Q02) is not as far. This can be interpreted as the largest effect size of DIF being present in item PM918Q02, whereas the smallest effect size of DIF is found in PM924Q02.

Figure 6

DIF Results Using Raju's Method



When Table 9 is examined, the three items shown in Figure 6 also have *p*-values less than .05, indicating the presence of DIF. Additionally, these three items appear to have positive values. This implies that these three items benefit the female group while disadvantaging the male group.

Table 9

DIF Results Using Raju's Method

Item	χ^2	*p-value*
PM00FQ01	0.1072	.9146
PM903Q03	0.8823	.3776
PM918Q01	−1.4403	.1498
PM918Q02	4.5377	.0000
PM918Q05	−0.8162	.4144
PM923Q01	1.2567	.2089
PM923Q03	0.8211	.4116
PM923Q04	1.3500	.1770
PM924Q02	2.1072	.0351
PM995Q01	2.7943	.0052
PM995Q02	0.7404	.4591
PM995Q03	1.3844	.1662

3.5. DIF Method with CDM Approach

In this section, the results of DIF detection using the CDM approach are shown using the Wald and LRT tests. The results of DIF detection based on the Wald and LRT tests are presented in detail in Tables 10 and 11.

Upon inspecting Table 10, it becomes evident that only four items lack DIF among the detected items lacked DIF. This can be interpreted as based on the wald test, nine items are found to be identified as containing DIF (PM00FQ01, PM903Q03, PM918Q01, PM918Q05, PM923Q01, PM923Q04, PM924Q02, PM995Q01, PM995Q02, and PM995Q03). If associated with the Q-Matrix structure in Table 2, item PM00FQ01 is related to attributes N3, P3, and C1; item PM903Q03 is related to attributes N1, P2, and C2; item PM918Q01 is associated with attributes N4, P3, and C3; item PM918Q05 is associated with attributes N4, P2, and C3; item PM923Q01 is related to

Table 10
DIF Results Using the Wald Test Method

Item	*Wald statistic*	*df*	*p-value*	*adj. p-value*	*DIF*
PM00FQ01	72633.1260	8	.00	0	+
PM903Q03	80326830.3923	8	.00	0	+
PM918Q01	0.0000	8	.00	0	+
PM918Q02	0.0000	8	1	1	-
PM918Q05	0.0000	8	.00	0	+
PM923Q01	182252.8104	8	.00	0	+
PM923Q03	378105629.3112	8	.581	1	-
PM923Q04	1535100739.7061	8	.00	0	+
PM924Q02	247483938448.5677	8	.00	0	+
PM995Q01	1949.3888	8	.00	0	+
PM995Q02	315791.1869	8	.00	0	+
PM995Q03	114.8911	8	.00	0	+

attributes N3, P2, and C4; item PM923Q04 is related to attributes N1, P1, and C4; item PM924Q02 is related to attributes N3, P1, and C1; item PM995Q01 is associated with attributes N2, P2, and C4; item PM995Q02 is associated with attributes N2, P1, and C4; and item PM995Q03 is related to attributes N3, P1, and C4.

Table 11
DIF Results Using the LRT Method

Item	*LR statistic*	*df*	*p-value*	*adj. p-value*	*DIF*
PM00FQ01	−0.9275	8	.9988	1	-
PM903Q03	−0.2885	8	1	1	-
PM918Q01	−0.8249	8	.0000	.0000	+
PM918Q02	−0.8249	8	1	1	-
PM918Q05	−0.8249	8	1	1	-
PM923Q01	9.5910	8	1	1	-
PM923Q03	27.1671	8	.0000	.0000	+
PM923Q04	30.2746	8	.0853	.7676	-
PM924Q02	−6.2723	8	.0000	1	-
PM995Q01	−51.1132	8	.9801	1	-
PM995Q02	2.6095	8	.9436	1	-
PM995Q03	−35.6247	8	.0000	.0003	+

Upon analyzing Table 11, it becomes evident that only three specific items (PM918Q01, PM923Q03, and PM995Q03) have been found to exhibit DIF, as determined through the application of the LRT technique. Of these three items, if associated with the Q-Matrix structure in Table 2, PM918Q01 is related to attributes N4, P3, and C3; PM923Q03 is related to attributes N2, P2, and C4; and PM995Q03 is related to attributes N3, P1, and C4.

To comprehend the nature of items with DIF in the CDM approach, it is necessary to investigate the prevalence and attribute probabilities based on the gender variable, as presented in Table 12 and Table 13. The prevalence of attributes is estimated by adding the probabilities of all pertinent latent classes. In the Indonesian sample, there were 2048 latent classes for 11 attributes. According to the information in Table 12, the easiest attribute for female students is C3 (societal). Approximately 69% of female students possessed this attribute. The most challenging attribute for female students is C4 (scientific), as only 13% of female students possess this attribute. For male students, the easiest attribute is C3, the same as for female students. Approximately 71% of male students possess this attribute. Meanwhile, the most difficult attribute for male students was C4, as only 30% of male students possessed this attribute. Although the easiest and most difficult attributes are the same for female and male students, male students tend to have a slightly higher mastery of attributes C3 and C4, with a difference of 2% and 7%, respectively. In contrast, it is clear that female students generally performed well in the remaining eight attributes, with the exception of attributes N3 (Quantity), P1 (Mathematization), and P2 (Mathematical operation).

Table 12
Mastery of Attributes based on Gender

Item	*Female*	*Male*
N1	0.48	0.46
N2	0.54	0.41
N3	0.24	0.35
N4	0.58	0.54
P1	0.48	0.56
P2	0.41	0.68
P3	0.61	0.59
C1	0.59	0.51
C2	0.50	0.46
C3	0.69	0.71
C4	0.13	0.30

Table 13
Attribute Profile

Latent Class	*Female*	*Male*
00000000000	0.00	0.00
01000010000	0.01	0.00
11000010000	0.01	0.00
11000010010	0.02	0.00
11000010110	0.02	0.00
11001010010	0.00	0.01
11010001110	0.01	0.00
11001010110	0.00	0.01
10111111100	0.01	0.00
10110111110	0.00	0.01
11111110110	0.00	0.01
11011111110	0.01	0.00
10111111110	0.01	0.00
11111111011	0.00	0.01
11111110111	0.00	0.01
10111111111	0.00	0.01
11111111110	0.00	0.02
11111111111	0.01	0.01

When Table 13 is examined, it is found that 0% of female and male students are in the latent class "00000000000." This can be interpreted as no students having mastered any of the 11 attributes. On the other hand, 1% of female and male students were in the latent class "11111111111," representing a mastery level of all attributes.

3.6. Comparison of DIF Methods

In this section, the results of the analysis based on all the methods are presented and compared, as listed in Table 14. According to the information in Table 14, when comparing the CTT method with the MH and LR techniques, both techniques show DIF for the same items (six items), namely PM918Q01, PM918Q02, PM918Q05, PM923Q01, PM923Q03, and PM924Q02. When comparing the IRT method with Lord's χ^2 technique (five items with DIF) and Raju's Unsigned Area Measures (three items with DIF), both techniques show DIF for the same items, namely PM918Q02 and PM995Q01. Both the CTT and IRT methods showed DIF for item PM918Q02. When investigating the CDM method, the Wald test shows DIF for ten items (PM00FQ01, PM903Q03, PM918Q01, PM918Q05, PM923Q01, PM923Q04, PM924Q02, PM995Q01, PM995Q02, and PM995Q03), whereas LRT only shows DIF for three items (PM918Q01, PM923Q03, and PM995Q03). Both the Wald test and LRT showed DIF for items PM918Q01 and PM995Q03. Overall, three items have been consistently identified as DIF in the CTT, IRT, and CDM methods, namely PM923Q01, PM923Q03, and PM924Q02.

Table 14

Comparison of DIF Results based on Different Methods

Item	*CTT Method*			*IRT Method*			*CDM Method*		
	MH	*LR*	*DIF*	*Lord*	*Raju*	*DIF*	*Wald*	*Lord*	*DIF*
PM00FQ01	-	-	0/2	-	-	0/2	+	-	1/2
PM903Q03	-	-	0/2	-	-	0/2	+	-	1/2
PM918Q01	+	+	2/2	-	-	0/2	+	+	0/2
PM918Q02	+	+	2/2	+	+	2/2	-	-	0/2
PM918Q05	+	+	2/2	-	-	0/2	+	-	0/2
PM923Q01	+	+	2/2	+	-	1/2	+	-	1/2
PM923Q03	+	+	2/2	+	-	1/2	-	+	1/2
PM923Q04	-	-	0/2	-	-	0/2	+	-	1/2
PM924Q02	+	+	2/2	-	+	1/2	+	-	1/2
PM995Q01	-	-	0/2	+	+	2/2	+	-	1/2
PM995Q02	-	-	0/2	-	-	0/2	+	-	1/2
PM995Q03	-	-	0/2	+	-	1/2	+	+	1/2

4. Discussion

This study aims to assess the fairness of mathematical literacy tests from a gender perspective using three DIF analysis approaches, CDM, CTT, and IRT, and compare the results of the three approaches to examine the compatibility between these approaches in identifying DIF effects. This study also answers various questions in the field of psychometrics, especially the issue of the accuracy of CDM use. It is important to answer these questions, considering the increasing need for large-scale assessment data using a CDM approach. Therefore, investigating the invariance of item parameters across different groups is important to ensure the appropriate use of the CDM. In this context, DIF analysis provides a suitable solution for investigating the validity of score interpretation. Terzi and Sen (2019) emphasize that attaching significance to the outcomes of large-scale evaluations without assessing the accuracy of score interpretation will not yield the anticipated advantages or produce the intended effect on policies. The significance of ensuring validity when conducting CDM analysis with substantial datasets should be highlighted in the literature, as variations in test language, intercultural differences, and demographic factors such as gender can result in alterations in student performance.

Twelve items from PISA 2012 related to mathematical literacy were analyzed using the CTT, IRT, and CDM approaches. When comparing the MH and LR methods based on CTT, DIF was found in the same items for both methods. Lord's x^2 method and Raju's Unsigned Area Measures method based on IRT were compared, and it was discovered that there was a DIF in two identical items, specifically PM918Q02 and PM995Q01. Only one item showed consistent DIF in all four methods, namely, PM918Q02. The CTT and IRT methods yielded identical results, except for Raju's Unsigned Area Measures method. This research is in line with previous studies that stated that both CTT and IRT mostly yield identical or consistent results (Eren et al., 2023). When comparing the DIF method based on CDM, the Wald test method detected DIF in ten items, whereas the LRT method was only able to detect three items. The two items detected through the LRT method were also detected using the Wald test method (PM918Q01 and PM995Q03). DIF was also identified in six other items using the Wald test. The results of this study contradict those of previous research, which indicates that the Wald test has the lowest DIF items and is different from the LRT method (Eren et al., 2023). These results align with the findings of Hou et al. (2020), who showed that the Wald test can detect DIF in most items and tends to be consistent with the findings of the LRT method.

Based on the prevalence and probability of attributes, the difference between groups of female and male students in items containing Differential Item Functioning based on the CDM method can be explained. For example, in item PM923Q03, male students tended to master all the attributes underlying the item more than female students (N2 = spatial and geometric concepts, P2 = mathematical operations, and C4 = scientific context). Similarly, in item PM923Q04, male students tended to master all the attributes underlying the item more than female students (N1 = change and relational concepts, P1 = mathematization process, and C4 = scientific context). DIF items on the content of change and relation and geomancy (space and shape) were also found in previous studies (Abedalaziz, 2010; Else-Quest et al., 2010; Shanmugam, 2018), and this was attributed to the visuospatial abilities of male students (Sanchis-Segura et al., 2018). In contrast, female students prefer algebra and probability (Abedalaziz, 2010; Shanmugam, 2018). Likewise, DIF items that require mathematical operations and mathematization skills are highly preferred by male students, while female students prefer to solve routine items (Abedalaziz, 2010; Kaiser & Zhu, 2022; Shanmugam, 2018). For DIF items using science and technology contexts, this has also been found in previous research, and is attributed to gender stereotypes formed through social and cultural contexts (Kaiser & Zhu, 2022; Niu, 2022; OECD, 2023). The lack of interest and representation of female students in science and STEM professional occupations makes it difficult to complete mathematical tasks in these contexts (OECD, 2023). When viewed in the PISA 2012 released items, both DIF items are identical to male students whose context is related to "SHAILING SHIPS" by utilizing kites to run ships to replace the role of engines that require a lot of diesels. According to Ong et al. (2015), such illustrations are highly exaggerated and may interfere with the validity of the score interpretation, as they do not belong to the construct being measured.

Overall, the findings of this research indicate that methods not based on the CDM model show fewer instances of Differential Item Functioning (DIF) compared to the Wald test, and more instances of DIF compared to the LRT method. These research findings are consistent with previous studies that show that the LRT method has fewer items with DIF compared to methods that are not based on CDM (Eren et al., 2023; Mehrazmay et al., 2021). There are two possible explanations for these differences. First, the test utilized in this research was not devised according to the cognitive diagnostic modeling framework (Ravand & Baghaei, 2020). Consequently, the test's psychometric properties may not have been entirely met, such as ensuring appropriate test development qualifications and defining the Q-matrix (Gierl et al., 2010). The retrofitting approach, which is not based on CDM, is commonly employed in large-scale assessment data, as has been noted by previous researchers. This is due to the fact that creating and implementing CDM-based tests can be difficult, particularly when it comes to ensuring the validity of the Q-Matrix, which must take into account various negative scenarios (Gierl et al., 2010; Li et al., 2020; Wang et al., 2018). The difficulty in defining the Q-Matrix greatly affects DIF findings in the MH method, Wald test, and LRT (Svetina et al., 2017). Second, regarding the sample size and discriminant indices. The number of items containing DIF tends to be inconsistent when different sample sizes are investigated and further compounded by low-item discriminant indices (Ma et al., 2021). According to Liu et al. (2019), the power of both the MH and LRT methods diminishes as the number of items exhibiting DIF grows. This can be interpreted as CDM-based methods such as LRT being more reliable than CTT-based methods such as MH.

5. Conclusion

This study contributes to the current research by focusing on ensuring fairness in testing and the validity of score interpretation. Methodologically, this research contributes by providing insights into DIF analysis using various methods from the CTT, IRT, and CDM approaches. In terms of findings, this research complements previous literature that predominantly focuses on the CTT

(MH), IRT (Raju), and CDM (Wald Tst) approaches and pays less attention to the suitability of the newer CDM-based approaches (Wald test and LRT) compared to traditional methods (CTT and IRT). The outcomes of this study reveal that of the 12 items assessed, there are variations in the conclusions drawn between the CTT, IRT, and CDM methods. The Raju Unsigned Area Measures method in the IRT and the Wald Test in the CDM approach revealed the item with the highest DIF, whereas the LRT method of the CDM approach identified the item with the lowest DIF. Furthermore, PM923Q01, PM923Q03, and PM924Q02 were identified as DIF items in all three methods: CTT, IRT, and CDM. Items PM923Q01 and PM923Q03 favor the group of male students, while item PM924Q02 favors the group of female students.

This study performed a DIF analysis to evaluate the psychometric properties of the test using the CDM framework rather than examining its source of bias. Future research can delve deeper by making more specific evaluations regarding item bias in terms of test structure, scope, and subgroups. In addition, the DIF analysis in this study was conducted using only gender variables and did not involve expert groups in validating the items that were considered biased. For this reason, future research could use different variables and involve expert groups to validate items that are considered biased based on the results of DIF analysis using the CTT, IRT, and CDM approaches.

Author contributions: All the authors contributed significantly to the conceptualization, analysis, and writing of this paper.

Availability of data and materials: The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Declaration of interest: No conflict of interest is declared by authors.

Ethics declaration: Since the data used in this study are publicly available, no ethical approval is required.

Funding: The financial support from the Indonesian Endowment Fund for Education (Lembaga Pengelola Dana Pendidikan-LPDP) of the Indonesian Ministry of Finance is gratefully acknowledged.

References

Abedalaziz, N. (2010). A gender- related differential item functioning of mathematics test items. *The International Journal of Educational and Psychological Assessment, 5,* 101-116.

Akbay, L. (2021). Impact of retrofitting and item ordering on DIF. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 212–225. https://doi.org/10.21031/epod.886920

Başman, M., & Kutlu, Ö. (2020). Identification of differential item functioning on mathematics achievement according to the interactions of gender and affective characteristics by Rasch Tree Method. *International Journal of Progressive Education, 16*(2), 205–217. https://doi.org/10.29329/ijpe.2020.241.14

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software, 48*(6). https://doi.org/10.18637/jss.v048.i06

De La Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*(2), 253–273. https://doi.org/10.1007/s11336-015-9467-8

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136*(1), 103–127. https://doi.org/10.1037/a0018053

Eren, B., Gündüz, T., & Tan, Ş. (2023). Comparison of methods used in detection of DIF in cognitive diagnostic models with traditional methods: applications in TIMSS 2011. *Journal of Measurement and Evaluation in Education and Psychology, 14*(1), 76-94. https://doi.org/10.21031/epod.1218144

Freudenthal, H. (1972). *Mathematics as an educational task.* Springer. https://doi.org/10.1007/978-94-010-2903-2

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R Package CDM for cognitive diagnosis models. *Journal of Statistical Software, 74*(2), 1-24. https://doi.org/10.18637/jss.v074.i02

Gierl, M. J., Alves, C., & Majeau, R. T. (2010). Using the attribute hierarchy method to make diagnostic inferences about examinees' knowledge and skills in mathematics: an operational implementation of cognitive diagnostic assessment. *International Journal of Testing, 10*(4), 318–341. https://doi.org/10.1080/15305058.2010.509554

Hou, L., La Torre, J. D., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: application of the wald test to investigate DIF in the DINA Model: Applying Wald test to investigate DIF in DINA model. *Journal of Educational Measurement, 51*(1), 98–125. https://doi.org/10.1111/jedm.12036

Hou, L., Terzï, R., & De La Torre, J. (2020). Wald test formulations in DIF detection of CDM data with the proportional reasoning test. *International Journal of Assessment Tools in Education, 7*(2), 145–158. https://doi.org/10.21449/ijate.689752

Kaiser, G., & Zhu, Y. (2022). Gender differences in mathematics achievement: A secondary analysis of Programme for International Student Assessment data from Shanghai. *Asian Journal for Mathematics Education, 1*(1), 115–130. https://doi.org/10.1177/27527263221091373

Kang, C., Yang, Y., & Zeng, P. (2019). Q-matrix refinement based on item fit statistic RMSEA. *Applied Psychological Measurement, 43*(7), 527–542. https://doi.org/10.1177/0146621618813104

Li, C., Ma, C., & Xu, G. (2020). *Learning large Q-matrix by restricted Boltzmann machines* (15424). arXiv. https://doi.org/10.48550/ARXIV.2006.15424

Li, T., & Traynor, A. (2022). The use of cognitive diagnostic modeling in the assessment of computational thinking. *AERA Open*, *8,* 1256. https://doi.org/10.1177/23328584221081256

Liu, Y., Yin, H., Xin, T., Shao, L., & Yuan, L. (2019). A comparison of differential item functioning detection methods in cognitive diagnostic models. *Frontiers in Psychology, 10*, 1137. https://doi.org/10.3389/fpsyg.2019.01137

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Routledge. https://doi.org/10.4324/9780203056615

Ma, W., & De La Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software, 93*(14), 1-26. https://doi.org/10.18637/jss.v093.i14

Ma, W., Terzi, R., & De La Torre, J. (2021). Detecting differential item functioning using multiple-group cognitive diagnosis models. *Applied Psychological Measurement, 45*(1), 37–53. https://doi.org/10.1177/0146621620965745

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847–862. https://doi.org/10.3758/BRM.42.3.847

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*(4), 719-748. https://doi.org/10.1093/jnci/22.4.719

Mehrazmay, R., Ghonsooly, B., & De La Torre, J. (2021). Detecting differential item functioning using cognitive diagnosis models: applications of the Wald test and likelihood ratio test in a university entrance examination. *Applied Measurement in Education, 34*(4), 262–284. https://doi.org/10.1080/08957347.2021.1987906

Moradi, Y., Baradaran, H., & Khamseh, M. E. (2016). Psychometric properties of the Iranian version of the diabetes numeracy Test-15. *International Journal of Preventive Medicine, 7,* 43. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4809128/

Niu, T. (2022). The Impact of Gender Difference on Major Selections of Chinese College Students. In A. Holl, J. Chen, & G. Guan (Eds.), *Proceedings of the 2022 5th International Conference on Humanities Education and Social Sciences* (ICHESS 2022) (pp. 216–224). Atlantis Press SARL. https://doi.org/10.2991/978-2-494069-89-3_25

OECD. (2018). *PISA for development assessment and analytical framework: reading, mathematics and science.* OECD. https://doi.org/10.1787/9789264305274-en

OECD. (2023). *PISA 2022 Results (Volume I): The state of learning and equity in education.* OECD. https://doi.org/10.1787/53f23881-en

Ong, Y. M., Williams, J., & Lamprianou, I. (2015). Exploring crossing differential item functioning by gender in mathematics assessment. *International Journal of Testing, 15*(4), 337–355. https://doi.org/10.1080/15305058.2015.1057639

Paulsen, J., Svetina, D., Feng, Y., & Valdivia, M. (2020). Examining the impact of differential item functioning on classification accuracy in cognitive diagnostic models. *Applied Psychological Measurement, 44*(4), 267–281. https://doi.org/10.1177/0146621619858675

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*(4), 495–502. https://doi.org/10.1007/BF02294403

Ravand, H., & Baghaei, P. (2020). Diagnostic classification models: recent developments, practical issues, and prospects. *International Journal of Testing, 20*(1), 24–56. https://doi.org/10.1080/15305058.2019.1588278

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications.* Guilford Press.

Rutkowski, L., & Rutkowski, D. (2016). A call for a more measured approach to reporting and interpreting PISA results. *Educational Researcher, 45*(4), 252–257. https://doi.org/10.3102/0013189X16649961

Sanchis-Segura, C., Aguirre, N., Cruz-Gómez, Á. J., Solozano, N., & Forn, C. (2018). Do gender-related stereotypes affect spatial performance? Exploring when, how and to whom using a chronometric two-choice mental rotation task. *Frontiers in Psychology, 9,* 1261. https://doi.org/10.3389/fpsyg.2018.01261

Shanmugam, S. K. S. (2018). Determining Gender Differential Item Functioning for Mathematics in Coeducational School Culture. *Malaysian Journal of Learning and Instruction, 15*(2), 83–109. https://doi.org/10.32890/mjli2018.15.2.4

Svetina, D., Dai, S., & Wang, X. (2017). Use of cognitive diagnostic model to study differential item functioning in accommodations. *Behaviormetrika, 44*(2), 313–349. https://doi.org/10.1007/s41237-017-0021-0

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Terzi, R., & Sen, S. (2019). A nondiagnostic assessment for diagnostic purposes: q-matrix validation and item-based model fit evaluation for the TIMSS 2011 assessment. *Sage Open, 9*(1), 215824401983268. https://doi.org/10.1177/2158244019832684

Wang, W., Song, L., Ding, S., Meng, Y., Cao, C., & Jie, Y. (2018). An EM-Based method for Q-matrix validation. *Applied Psychological Measurement, 42*(6), 446–459. https://doi.org/10.1177/0146621617752991

Wu, X., Wu, R., Chang, H.-H., Kong, Q., & Zhang, Y. (2020). International comparative study on pisa mathematics achievement test based on cognitive diagnostic models. *Frontiers in Psychology, 11*, 2230. https://doi.org/10.3389/fpsyg.2020.02230

Yildirim, O. (2019). Detecting gender differences in PISA 2012 mathematics test with differential item functioning. *International Education Studies, 12*(8), 59. https://doi.org/10.5539/ies.v12n8p59