REID (Research and Evaluation in Education), 10(1), 2024, 99-113
Available online at: http://journal.uny.ac.id/index.php/reid

Differential Item Functioning of the region-based national examination equipment

Adi Setiawan[1]*; Gulzhaina Kuralbaevna Kassymova[2]; Vianney Mbazumutima[3]; Anggit Reviana Dewi Agustyani[4]
[1]PT Batamindo Green Farm, Indonesia
[2]Abai Kazakh National Pedagogical University, Kazakhstan
[3]African Institute for Mathematical Sciences, Cameroon
[4]Umeå Mathematics Education Research Centre (UMERC), Sweden
*Corresponding Author. E-mail: adsetwan@gmail.com

ARTICLE INFO

Article History
Submitted:
16 May 2024
Revised:
28 June 2024
Accepted:
29 June 2024

Keywords
comparison of DIF detection methods; differential items functioning; unidimensional IRT

ABSTRACT

This research aims to detect Differential Item Functioning (DIF) in the 2014/2015 National Examination Questions in mathematics of junior high schools and equivalent-level schools in the Yogyakarta region as a reference group and the South Kalimantan region as a focus group using the Likelihood Ratio Test (LRT) method, Area Measure Raju, and Lord. A sensitivity analysis was conducted to determine the most sensitive method. The data consisted of 5,465 National Examination papers of the students from the two regions who worked on type A questions. A sample of 1,000 exam papers for each region was established using the simple random sampling (SRS) technique, which was conducted to avoid the effect of sample size. The research results showed that by using the LRT method, the researchers found 36 items had significant DIF detection, 32 items were significant for Raju Area, and all items had significant DIF detection using Lord. Lord Method is the most sensitive method because it can detect most DIF items.

Scan Me:

To cite this article (in APA style):
Setiawan, A., Kassymova, G., Mbazumutima, V., & Agustyani, A. (2024). Differential Item Functioning of the region-based national examination equipment. *REID (Research and Evaluation in Education), 10*(1), 99-113. doi:https://doi.org/10.21831/reid.v10i1.73270

INTRODUCTION

The education assessment system in Indonesia has undergone significant transformation over several decades. Starting from the Final Examination (1950-1964), this system was then transformed into the National Examination or *Ujian Nasional* (UN) in 2005. For 15 years, the UN became the main measuring tool for assessing the quality of national, regional, and individual education and the basis for selection to further education (Yamin & Syahrir, 2020). The National Examination could determine a student's graduation. Implemented in elementary, middle, and high schools, the National Examination has been a benchmark for student achievement for over a decade. However, in 2020, the National Examination was officially abolished and replaced with a new assessment system, namely, the National Assessment or *Asesmen Nasional* (AN) (Azis, 2015). This decision was made because the National Examination was considered less effective in measuring student learning achievement and character (Hadi et al., 2020). The abolition of the

ISSN: 2460-6995 (Online)

National Examination marks a new era in educational assessment in Indonesia (Alfarizi, 2019). AN is designed to provide a more comprehensive picture of the quality of education, with a focus on measuring literacy, numeracy, and character (Jusmirad et al., 2023; Zukmadini et al., 2021). This system is expected to produce more comprehensive data to map the quality of education at various levels and regions and become the basis for developing more targeted education policies (Hidajad, 2019).

To replace the UN, there is a need for a new test device that is valid and consistent (Ihsan, 2016). Validity refers to the ability of a test to measure what it is supposed to measure, while consistency ensures that measurement results are stable and not influenced by external factors (Gaberson, 1997; Sinha et al., 2013). One indicator of an invalid and consistent test is Differential Items Functioning (DIF) (French et al., 2019; Kane, 2013). DIF occurs when the performance of test takers from a certain group (for example, gender, region, ethnicity) differs significantly on certain items compared to the reference group (Cho et al., 2016; Delgado et al., 2018; Kane, 2013; Retnawati, 2013). To detect DIF, test participants are divided into two groups: a focus group and a reference group. The focus group consists of test takers who are suspected of being disadvantaged by certain test items compared to the reference group (Desjardins & Bulut, 2017). This grouping can be based on various factors, such as gender, region, and ethnicity. Test items that show DIF indicate unfairness in the test (Effiom, 2021). These items favor certain groups and cannot objectively measure test takers' ability. Therefore, detecting and removing DIF items from new test sets is important to ensure fair and accurate assessment (Effiom, 2021).

Much research has been carried out regarding the DIF in various contexts. Akour et al. (2015) detected DIF in the 2006 Program for International Student Assessment (PISA) data using the Net and Global Differential Items Functioning method and found that five of the six items contained DIF. Patricia and Araújo (2012), in their research on DIF in PISA 2009 reading ability items based on student immigrant status, classified DIF items based on question format, text format, aspect, type, and text situation. Yildirim (2019) detected DIF in the 2012 PISA mathematics test based on gender and found three questions containing DIF. Meanwhile, Zampetakis et al. (2017), in their research with survey data measuring the variables entrepreneurial intuition, attitudes towards entrepreneurship, subjective norms, and perceived behavioral control, found DIF in the sixth item, "taking time to learn about starting a company" on the latent variable entrepreneurial intuition.

Various methods are used to detect DIF, including Factor Analysis, Mantel-Haenszel, Log-linear model, chi-square, ANOVA, testing item differentiation using point-biserial and partial correlation, testing item difficulty levels using various theoretical transformations, Likelihood Ratio Test, Lord Chi-Square test, and Raju Area Measure. These methods are generally divided into two categories: parametric DIF (or DIF based on item response theory) and nonparametric DIF. Parametric DIF uses item and individual ability level parameters to detect DIF, while nonparametric DIF uses only raw items and test scores. Desjardins and Bulut (2017) state that parametric DIF is better than nonparametric DIF.

Several methods are commonly used in parametric DIF detection, including the Likelihood Ratio Test (LRT) and Raju Area Measure. LRT was developed by Thissen et al. (1986) and can detect significant differences in item responses between two groups. The LRT does not require a variance-covariance matrix parameter estimator, so it is theoretically preferable to the Raju Area Measure (Thissen et al., 1988). Raju's Broad Measure, developed and refined by Raju (1990), compares the function of item characteristics, namely the Item Characteristics Curve (ICC) (Hambleton et al., 1991). Various DIF methods produce different estimates, with the intent of finding the best method. Sensitivity analysis can be used to determine the most sensitive method. This is done by identifying the method that detects the most DIF. The method that detects the most DIF in the data used is considered the most sensitive method. Research related to sensitivity analysis based on the number of DIF items has been carried out.

Mathematics is a scientific discipline that emphasises organized thought patterns, logical proof, and clear and concise symbolic representation (James et al., 1959). Introduced from an early

age, mathematics became the basis for studying other sciences. Mathematics material is neatly structured, so the knowledge and understanding obtained previously become the foundation for studying subsequent material. Mathematics, as one of the basic subjects in education, makes Mathematical ability an important factor in measuring student achievement. Student achievement can be measured through various measuring tools, including the UN.

Based on the results of the UN in mathematics at the junior high school and equivalent-level schools in 2015-2019 (Center for Educational Assessment, 2020), the Yogyakarta region consistently shows superior performance compared to the national average score. This can be seen in Figure 1, where the Yogyakarta area has always been above the national average line for five consecutive years. On the other hand, several regions are still below the national average, such as the South Kalimantan region.

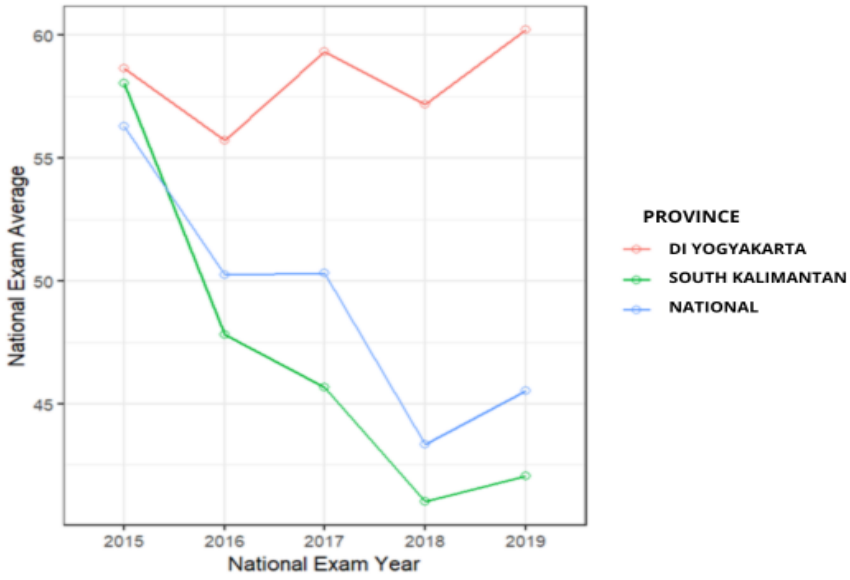


Figure 1. Average National Examination Score for Middle School in Mathematics

Figure 1 shows that the Yogyakarta region is consistently above the national average score for the National Examination Mathematics for junior high school and equivalent-level schools during the 2015-2019 period, while the South Kalimantan region is always below it. In 2015, both regions showed almost the same scores. The DKI Jakarta region, with an average score of 71.19, has the highest performance. Differences in National Examination results can be influenced by various factors, one of which is the fairness of the test equipment (Leiner et al., 2018; Siegrist et al., 2012). Therefore, it is necessary to test the test equipment (National Examination) by region to determine whether it is fair for all groups. A fair test set is characterized by the absence of questions containing DIF.

DIF detection is done by dividing test participants into two groups: a reference group and a focus group. The Yogyakarta region was chosen as a reference group because of its good quality of education and because it is a reference for other regions in Indonesia. The Yogyakarta region also has the highest teacher competency in Indonesia (Turang, 2017; Sitepu & Rahmawati, 2022), contributing to high student abilities and good National Examination results. In the 2014/2015 Junior high school and equivalent-level schools National Examination Mathematics, the average score for the Yogyakarta region was 58.66, and only South Kalimantan had a score close to that, namely 58.05 (Center for Educational Assessment, 2020). This was the reason for selecting South Kalimantan as the focus group. DIF detection based on region has been carried out in various studies. Whynes et al. (2013) detected DIF on the health-related quality of life instrument (EQ-5D) with acute stroke clinical trial data (ISRCTN 99414122). The results show that the average service index score in the United Kingdom is significantly higher than in Asia and elsewhere. Another

study by Huang et al. (2016) detected DIF by region (United States-Canada, Mainland China-Hong Kong China, and United States-Mainland China) in PISA data using the Multidimensional Random Coefficient Multinomial Logit Model (MRCMLM) method. The results showed that the number of easier questions to do are seven questions for American students, three questions for Canadian students, 19 questions for Mainland Chinese students, 17 questions for Hong Kong Chinese students, 31 questions for American students, and 29 questions for Mainland Chinese students.

Detecting DIF on National Examination test equipment based on region is very important because Indonesia's education quality is not evenly distributed. Fair testing tools for all students in Indonesia are needed to ensure fairness in achievement assessments. The results of this DIF detection will be used to evaluate the creation of a new test device, which is expected to be better and fairer. Thus, this research aims to detect DIF on National Examination test equipment using various methods, reveal the most sensitive method for detecting DIF on National Examination test equipment, and reveal which items containing DIF tend to be more beneficial for students in which areas.

METHOD

This research uses UN data from two regions: Yogyakarta and South Kalimantan. The research population consisted of 51,010 junior high school/equivalent students in the Yogyakarta area and 56,852 junior high school students in South Kalimantan (Center for Educational Assessment, 2020). The research sample was established randomly without replacement using the stratified random sampling method. Stratification was carried out based on region to avoid the influence of sample size (Scott et al., 2009). From each region, 1,000 students were taken to work on type A questions. Moreover, the research participants were anonymized to maintain confidentiality, and the data obtained were used only for research purposes.

The variables used in this research include region and the 2014/2015 National Examination (UN) questions in mathematics for junior high school/equivalent-level schools, from number 1 to 40. The regional variable identifies the school location of junior high school/equivalent-level students in the research sample. This study has two scores for the regional variable: the Yogyakarta region and the South Kalimantan region. Meanwhile, the question item variables, from numbers 1 to 40, each have one column in the research data. Each column contains a dichotomous score, where a score of 0 indicates an incorrect answer, and a score of 1 indicates a correct answer. Model fit analysis and assumption testing were used to select the best-fitting model for both data. Furthermore, DIF analysis was carried out to detect the presence of DIF in the test items in the data used. DIF analysis was carried out using the LRT, Raju, and Lord Methods. The method that detects most DIF is the most sensitive. The data analysis process was carried out using R Studio Open-Source software.

FINDINGS AND DISCUSSION

Findings

Model suitability and assumptions

In this study, the best model selection was based on the model with the most suitable items and the model with the smallest AIC, BIC, and AICc values, as shown in Table 1.

Table 1. Comparison of the Number of Suitable Test Items in Data for the Yogyakarta and South Kalimantan Regions

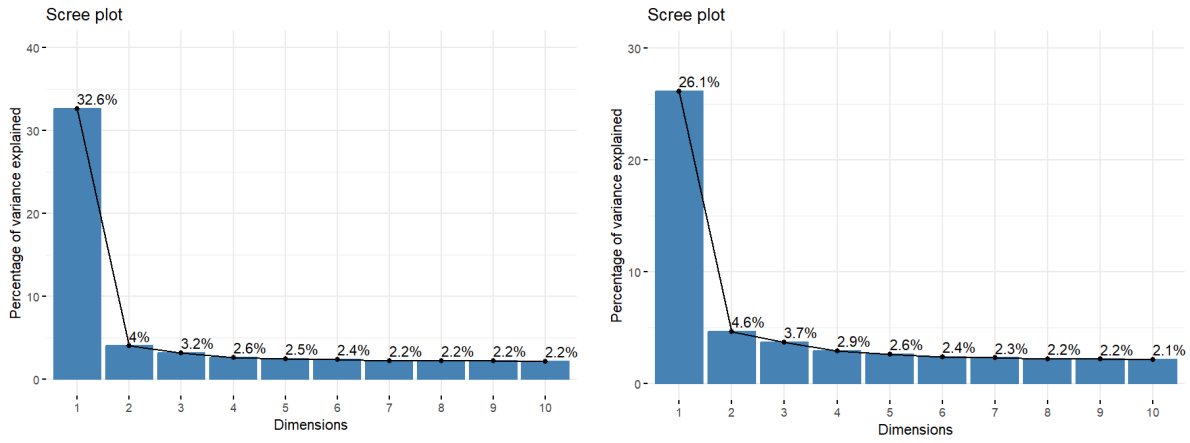
Rasch		1PL		2PL		3PL	
Yogyakarta	South Kalimantan	Yogyakarta	South Kalimantan	Yogyakarta	South Kalimantan	Yogyakarta	South Kalimantan
10	5	9	5	28	18	32	23

Table 1 shows that the 3PL model is the most suitable for data from the Yogyakarta and South Kalimantan regions because it has the largest number of suitable items. In addition, it is necessary to look at the goodness of fit for each model, using AIC, BIC, and AICc.

Table 2. The Goodness of Fit in Each Model in Each Data

Model	Region	AIC	BIC	AICc
Rasch	Yogyakarta	37691.85	37893.07	37695.45
1PL	Yogyakarta	37692.28	37893.50	37695.88
2PL	Yogyakarta	37025.18	37417..80	37039.28
3PL	**Yogyakarta**	**36375.80**	**36964.73**	**36408.84**
Rasch	South Kalimantan	47540.17	47741.39	47543.77
1PL	South Kalimantan	47540.17	47741.39	47543.77
2PL	South Kalimantan	46403.83	46796.45	46417..93
3PL	**South Kalimantan**	**45955.39**	**46544.32**	**45988.43**

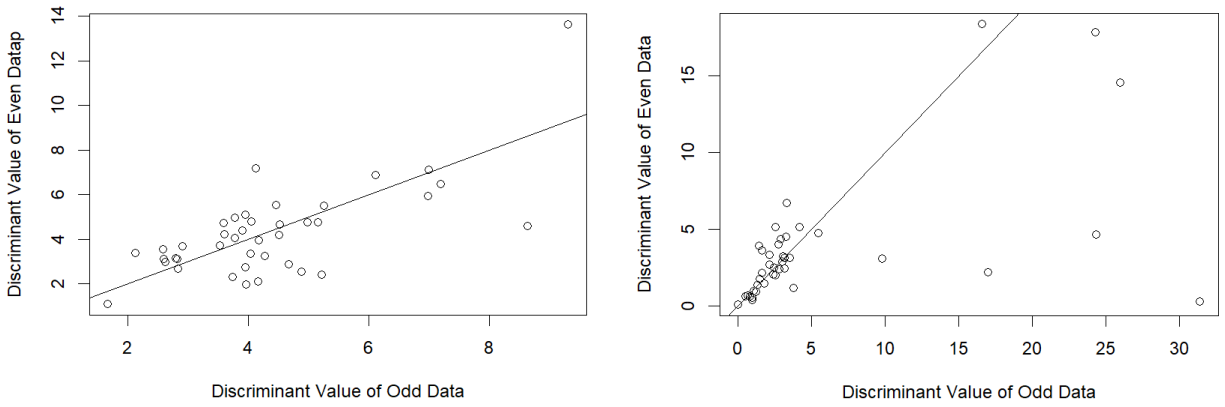
Table 2 shows that the 3PL model in the data for the Yogyakarta region and the South Kalimantan region has the lowest AIC, BIC, and AICc values, so it can be concluded that the best model for these two data is the 3PL model. Therefore, in the next analysis, the 3PL model was used. Based on testing the unidimensional assumption, factor analysis using 40 variables (test items), the factor analysis the eigenvalues were obtained, which were then plotted in an ordered manner (scree plot) because the ordered eigenvalues number 11 onwards tend to be the same (constant). Then, the 10 highest-ordered eigenvalues were displayed. In the scree plot, the amount of variance explained by the eigenvalues is displayed, and the data for the Yogyakarta Region can be seen in Figure 2a and for the South Kalimantan Region in Figure 2b.



(a) (b)

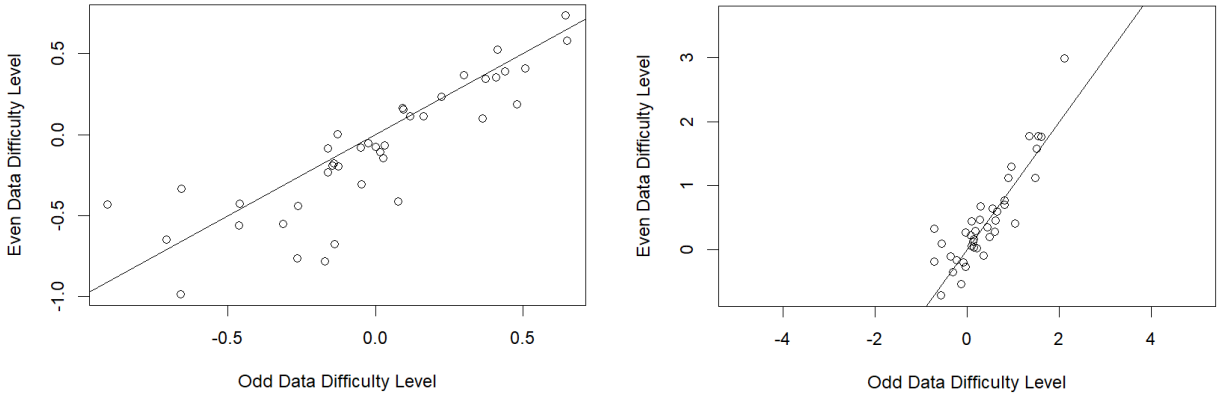
Figure 2. (a) Scree Plot on Yogyakarta Data, (b) Scree Plot on South Kalimantan Data

Figure 2a and Figure 2b show that one dimension is very dominant over the other dimensions, and there is an elbow point from dimension one to dimension two so that the unidimensional assumption is met. The local independence assumption is met because the unidimensionality assumption is met. Testing the assumption of parameter invariance was carried out by dividing student data into two groups: 500 student data in odd order and 500 student data in even order. After that, it was estimated using the 3PL model, obtaining parameter estimates a (discriminant), b (difficulty level), and g (Pseudo-Guessing). Figure 3a and Figure 3b are invariance plots of parameter a, Figure 4a and Figure 4b are invariance plots of parameter b, Figure 5a and Figure 5b are invariance plots of parameter g, and Figure 6a and Figure 6b are ability invariance plots (θ). The invariance assumption is met because the points follow a straight line, as seen in all the invariance plot figures.



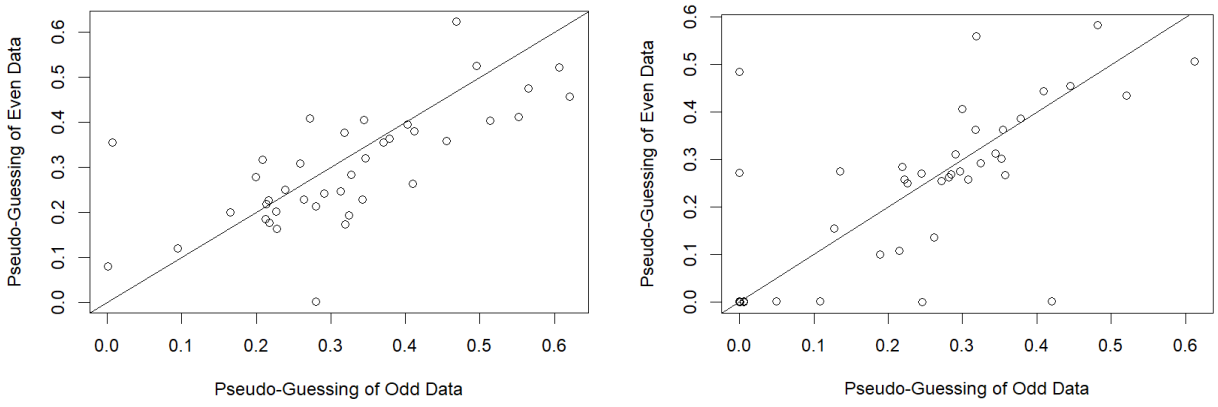
(a) (b)

Figure 3. Invariance of Discriminant Parameters in Data from (a) Yogyakarta Region and (b) South Kalimantan Region



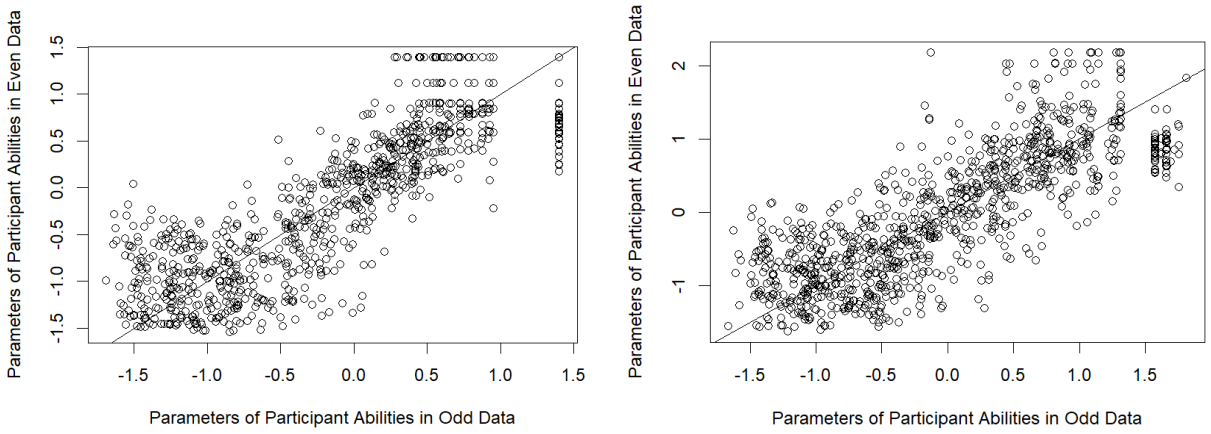
(a) (b)

Figure 4. Invariance of Difficulty Level Parameters in Data from (a) Yogyakarta Region and (b) South Kalimantan Region



(a) (b)

Figure 5. Invariance of Pseudo-Guessing Parameters in Data from (a) Yogyakarta Region and (b) South Kalimantan Region



(a) (b)

Figure 6. Invariance of competency parameters in data from (a) Yogyakarta Region and (b) South Kalimantan Region

Differential Items Functioning

In this study, the researchers used three methods to detect DIF: the Likelihood Ratio Test Method, the Raju Area Measure Method, and the Lord Method. The results of DIF detection using the Likelihood Ratio Test Method can be seen in Table 3. In LRT, the calculated values are compared with a table of degrees of freedom m, where m is the difference in the number of parameters estimated between the compact and augmented models. In this study, the compact model consisted of 39 questions, and the augmented model consisted of 40 questions. The number of parameters estimated in the compact model is 117 (three parameters in each item), and the number of parameters in the augmented model is 120 (three parameters in each item). Therefore, the degree of freedom m is equal to 3. Items can be said to have DIF detected if the statistic is greater than 7.81. The results of DIF detection using the Likelihood Ratio Test Method showed that 36 items had DIF detected; only items 8, 21, 26, and 34 were not detected by DIF.

Table 3. LRT Method DIF Detection Results

Test Item Number	Statistics	*P-Value*	Category	Test Item Number	Statistics	*P-Value*	Category
1	16.68	0.0008	DIF	21	4.76	0.1906	Not DIF
2	8.28	0.0405	DIF	22	9.45	0.0239	DIF
3	19.89	0.0002	DIF	23	66.16	0	DIF
4	37..84	0	DIF	24	8.79	0.0322	DIF
5	15.4	0.0015	DIF	25	11.34	0.01	DIF
6	31.02	0	DIF	26	7..81	0.0501	Not DIF
7	9.92	0.0193	DIF	27	399.88	0	DIF
8	5.24	0.1549	Not DIF	28	17..51	0.0006	DIF
9	124.19	0	DIF	29	144.86	0	DIF
10	29.27	0	DIF	30	16.96	0.0007	DIF
11	75.57	0	DIF	31	62.15	0	DIF
12	29.38	0	DIF	32	49.42	0	DIF
13	11.77	0.0082	DIF	33	83.4	0	DIF
14	146.4	0	DIF	34	2.12	0.5476	Not DIF
15	29.24	0	DIF	35	29.57	0	DIF
16	28.15	0	DIF	36	25.66	0	DIF
17	62.3	0	DIF	37	23.71	0	DIF
18	15.67	0.0013	DIF	38	21.85	0.0001	DIF
19	60.37	0	DIF	39	79.53	0	DIF
20	75.92	0	DIF	40	129.75	0	DIF

The results of DIF detection using the Lord Method are shown in Figure 7. Where items detected by DIF have a value of more than 5.99 (α=0.05), items outside the line are colored red, while items that do not contain DIF are shown in black. Figure 7 shows that all the items are red and outside the line. Therefore, DIF detection using the Lord Method concludes that significantly all items contain DIF.

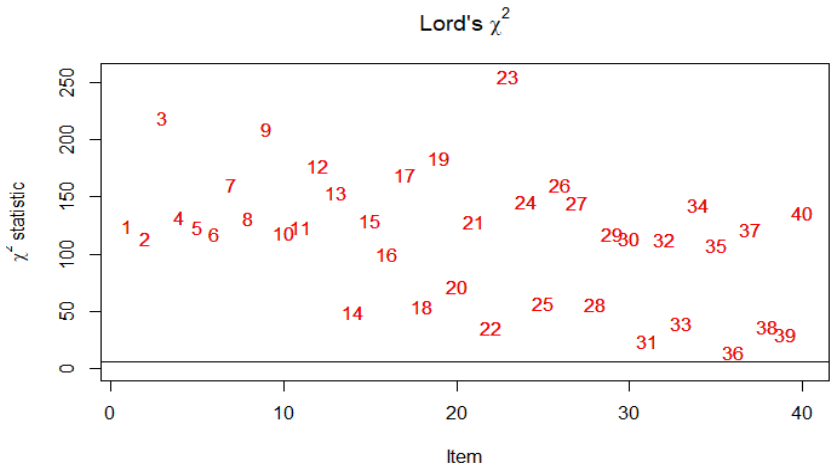


Figure 7. Result of Lord Method DIF Detection

The results of DIF detection using Raju Area Measures (SA method) can be seen in Figure 8. The grains detected by DIF are outside the parallel lines in the value range of -1.96 to 1.96. Items that are outside the line are colored red, while items that do not contain DIF are colored black. This shows that only test items number 2, 4, 6, 11, 22, 27, 35, and 38 do not contain DIF.

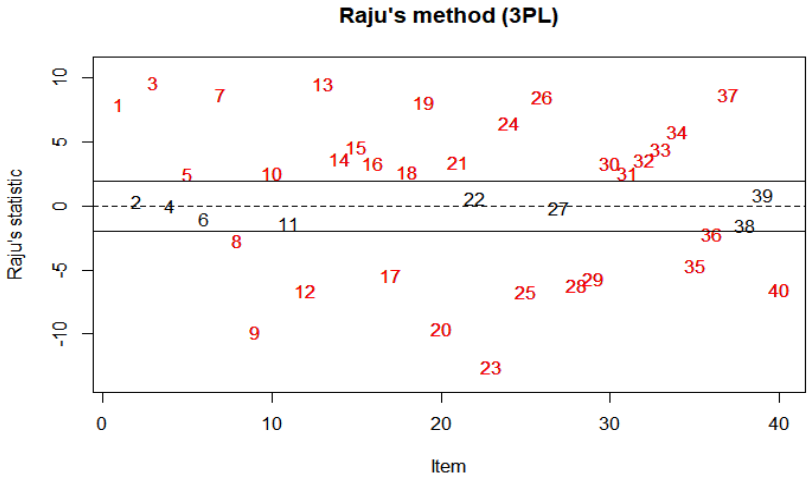


Figure 8. Result of Raju Method DIF Detection

DIF Sensitivity Level

In this research, DIF was analyzed using three methods, namely Likelihood Ratio Test, Raju, and Lord. The analysis results show that the Lord Method produces the highest number of DIF items, namely 40 items, followed by the Raju Method with 32 items and the Likelihood Ratio Test Method with 36 items. This shows that Lord Method is the most sensitive in detecting DIF in the 2015 Middle School National Examination mathematics data compared to the Yogyakarta and South Kalimantan regions. The percentage comparison of DIF detection results with various methods can be seen in Figure 9.

These findings indicate that the Lord Method can be used more effectively to identify items on the 2015 Middle School National Examination Mathematics, which shows differences in function between the reference and focus groups. This is important to ensure the fairness and validity of the National Middle School National Examination mathematics test so that the test results can more accurately reflect the abilities of students from various regions.

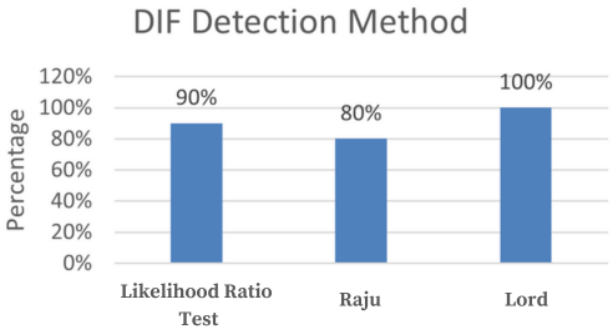


Figure 9. Comparison of DIF Detection Results with Various Methods

Probability of a Student's Correct Answer Based on Regions

Items that previously contained significant DIF were analyzed using ICC to see the probability of students answering correctly based on region, as shown in Figure 10 and Figure 11. Yogyakarta region is shown in a black line, while South Kalimantan region is shown in a red line.

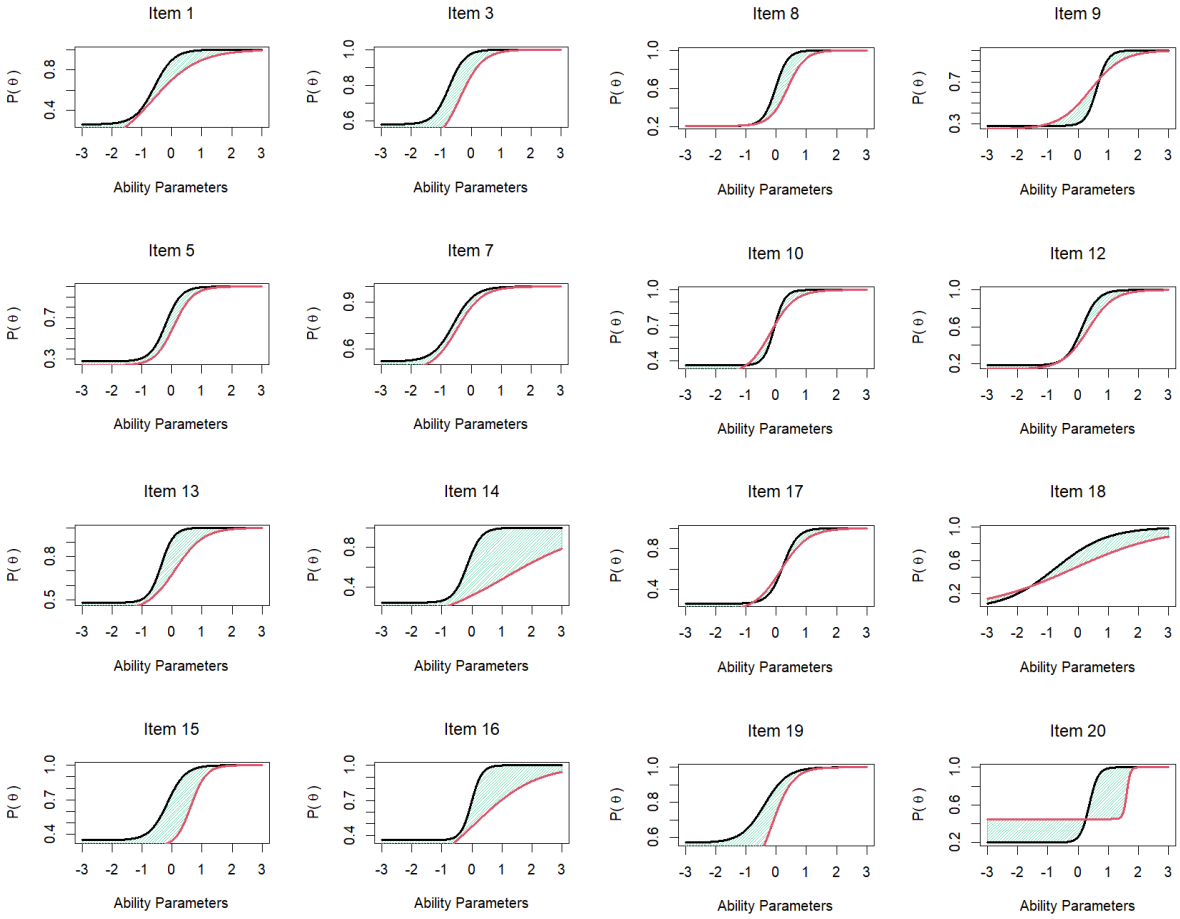


Figure 10. ICC between the Regions of Yogyakarta and South Kalimantan on DIF-Detected Items 1-20

Figure 10 shows that the students from the Yogyakarta Region have the opportunity to answer correctly items 1, 3, 5, 7, 8, 12, 13, 14, 15, 16, and 20. In items 9, 10, and 17, it can be seen that the interval is divided into three, and they can be classified into low, medium, and high ability. Students from the Yogyakarta Region with low and high abilities are superior and have a higher chance of answering questions correctly than students from the South Kalimantan Region. For Items 18 and 20, the interval is divided into two: low and high ability. Students from the South Kalimantan Region with a low ability level have a greater chance of answering questions correctly than students from the Yogyakarta Region. However, students from the Yogyakarta Region have a high chance of having a high ability level.

ICC for items 21 to 40 can be seen in Figure 11. Items 21, 24, 26, 28, 30, 32, 33, 34, and 36 show that the ability of students from the Yogyakarta Region is better in answering a test item correctly than students from the South Kalimantan Region. However, the opposite is shown in Items 29 and 40, where students from the South Kalimantan region are more likely to answer questions correctly at all levels of ability: low, medium, and high. In Items 23, 25, 31, 35, and 37, the interval is divided into two, which can be classified into low and high ability. Students from the Yogyakarta Region with high ability have a greater chance of answering questions correctly than those from the South Kalimantan Region. In contrast, students from the South Kalimantan Region only have a high chance of answering questions correctly at a low ability level.

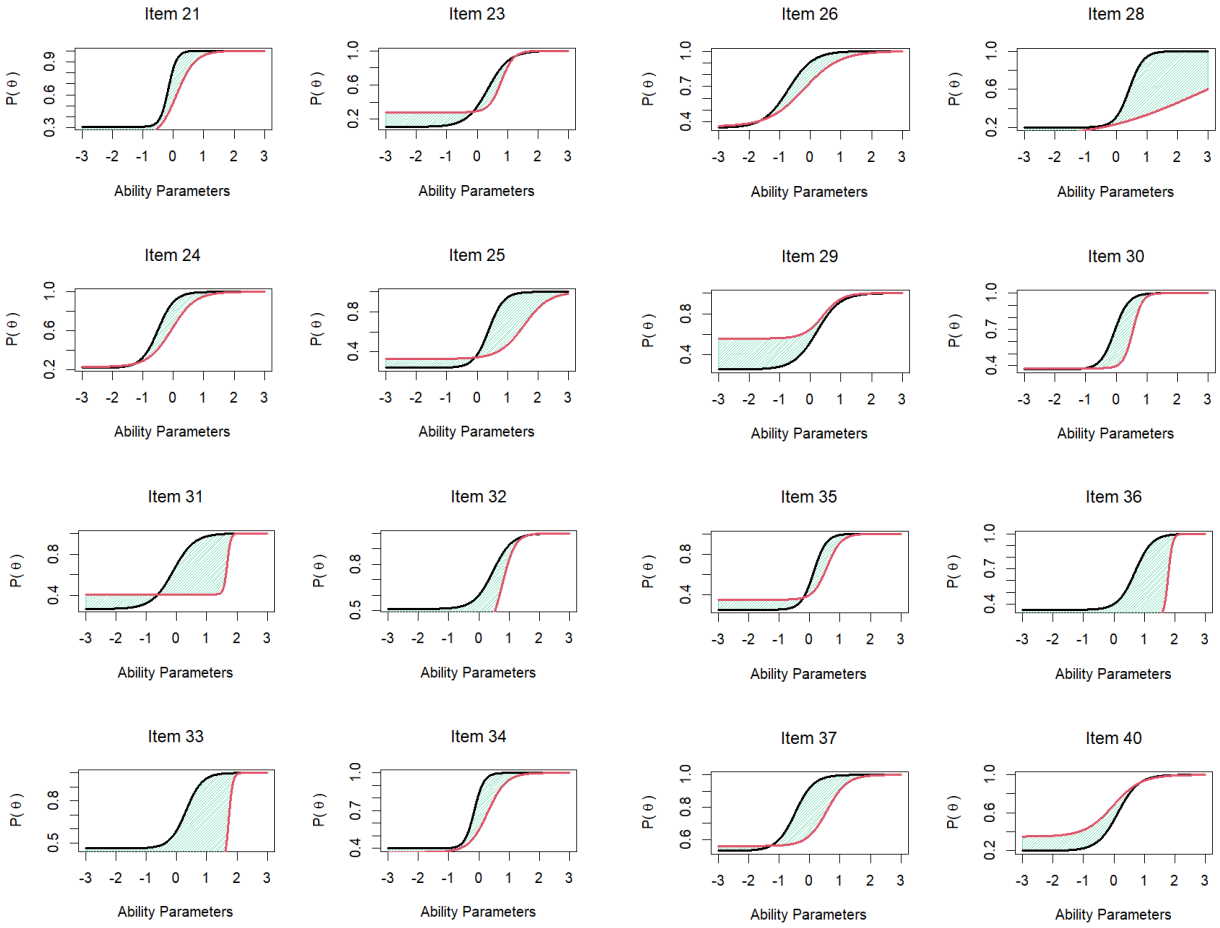


Figure 11. ICC between Yogyakarta Region and South Kalimantan Region in Items 21-40 Detected by DIF

Discussion

The results of the analysis show that there are still many items that contain DIF. In this study, the total number of items that significantly contained DIF in the 2014/2015 National Examination (UN) mathematics questions for junior high school/equivalent level, from numbers 1 to 40, was quite large. A total of 36 items that significantly contained DIF were detected using the Likelihood Ratio Test Method, 32 items that significantly contained DIF using the Raju Broad Measure Method, and 40 items or all items contained DIF through detection using the Lord Method. This result is quite worrying because the items detected are UN items. Previously, research on DIF conducted by Sudaryono (2017) also found the same thing in the national mathematics examination questions for senior high schools in Tangerang for the 2008/2009 academic year. Research by Hadi et al. (2021) also shows the same thing. DIF still detected many items on the 2014/2015 high school level mathematics national examination test using the Rasch method based on region. However, in contrast to research by Galli et al. (2011), which stated that only eight out of 30 items detected DIF based on the region using the Likelihood Ratio Test Method on a mathematics test device (not the National Examination) which was tested on participants who did not have a mathematics background. Even though there are differences in number, questions containing DIF should still be given special attention. Our findings show that the National Examination test equipment still has many shortcomings, especially in making questions. It is recommended that question-making should pay more attention to the socio-demographic conditions and abilities of test takers in each region so that the questions can be fair. It is also necessary to look more closely at the national scale test development process so that the test items do not contain DIF.

The analysis results also show that the method that produces the most items containing DIF is the Lord Method, which is the most sensitive. These results are in line with those found by Langer (2008) that the Wald test, a variation of the Lord Method, performed better in detecting DIF. In addition, Soysal and Koğar (2021) also showed the same thing, that the Lord Method was the most sensitive in detecting DIF, especially when considering the effect of item position. This was further supported by Sudaryono (2017), who found that the Lord Method outperformed other methods, including Mantel-Haenszel and Scheuneman's Chi-square, in detecting DIF. On the other hand, previous studies on DIF detection have identified various key factors that can affect the performance of detection methods. Uğurlu and Atar (2020) found that sample size and percentage of items with DIF significantly impact the performance of Multiple Indicators, Multiple Causes, and Logistic Regression methods. Similarly, Başman (2023) and Ukanda et al. (2019) both highlight the importance of sample size, DIF ratio, and test length in the efficacy of DIF detection methods, with the Logistic Regression and Mantel-Haenszel methods showing the lowest Type I error rates and the highest power levels. Berrío et al. (2019) further emphasize the need for model fit and the impact of sample size ratios on the power of the Difficulty Parameter Difference procedure. These findings underscore the importance of considering these factors in designing and implementing DIF detection methods. Therefore, DIF in the questions can be detected accurately. These results differ from research by Effendi (2011), which stated that the Likelihood Ratio Test Method was the most sensitive. In his research, which used National Examination data for high school level chemistry in 2008, 12 items had significant DIF based on gender using the LRT method, eight items based on the Raju Method, and seven items based on the Lord Method. The detection of DIF in all items in this study using the Lord Method also occurred in research conducted by Çelik and Özkan (2020). Çelik and Özkan (2020) research detected the presence of DIF in PISA 2015 data based on regions in Turkey using the Rasch model method. The results of their research showed that all items had significant DIF detected. According to Ozdemir and Alshamrani (2020), detecting and overcoming DIF is crucial to preventing biased assessments.

We also analysed the probability of students answering questions correctly based on their region of origin by using ICC on items containing DIF, detected using the Raju area measure method. The analysis results show that only two items are profitable for students from the South Kalimantan Region (focus group): Items 29 and 40. Other items are more profitable for students from the

Yogyakarta Region (reference group). The form of the questions in these two items can be seen in Figure 12 for item 29 and Figure 13 for item 40. These two images show that these questions have a distinctive characteristic: answering them requires only one step. For Item 29, participants can answer only one theory; for Item 40, participants can answer only one probability theory.

29. Observe the image!
Line AE is...
A. median
B. bisector
C. altitude
D. Axis

29. Perhatikan gambar!
Garis AE adalah ....
A. garis berat
B. garis bagi
C. garis tinggi
D. garis sumbu

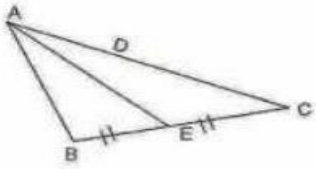


Figure 12. Test Item No. 29

40. Dalam kegiatan gerak jalan santai yang diikuti oleh 150 peserta, panitia menyediakan hadiah 3 buah sepeda. Peluang setiap peserta untuk mendapatkan hadiah adalah ....
A. 0,02
B. 0,03
C. 0,20
D. 0,30

40. In the leisure walking event attended by 150 participants, the committee provided 3 bicycles as prizes. The probability of each participant winning a prize is ....
A. 0.02
B. 0.03
C. 0.20
D. 0.30

Figure 13. Test Item No. 40

CONCLUSION

Based on the results of the analysis, it is concluded that of the 40 items analyzed, all of them have the potential to contain DIF. Analyzed by using the Likelihood Ratio Test Method, 36 items were detected to contain DIF; by using the Raju Area Measure Method, 32 items were detected, and the Lord Method showed that all items contained DIF. Thus, of the three methods used, the Lord Method is the most sensitive DIF detection method. Of the 40 questions, 38 favoured students from the Yogyakarta region, and only two favoured students from the South Kalimantan region. As a recommendation for future research, it is necessary to detect DIF by looking at other factors such as gender, school location, and other factors that might cause DIF in test items.

ACKNOWLEDGMENTS

The authors sincerely thank Heri Retnawati for aiding and guiding them in writing this article.

DISCLOSURE STATEMENT

The authors declare that there is no conflict of interest to disclose.

REFERENCES

Akour, M., Sabah, S., & Hammouri, H. (2015). Net and global differential item functioning in PISA polytomously scored science items. *Journal of Psychoeducational Assessment*, *33*(2), 166–176. https://doi.org/10.1177/0734282914541337

Alfarizi. (2019). Meningkatkan mutu pendidikan di Indonesia melalui MESUPPEN "Maksimalkan pendekatan supervisi pendidikan." *Tugas Kuliah Administrasi dan Supervisi Pendidikan Jurusan Matematika Universitas Negeri Padang*, 1–5. http://dx.doi.org/10.31227/osf.io/tmyz7

Azis, A. (2015). Conceptions and practices of assessment: A case of teachers representing improvement conception. *TEFLIN Journal - A Publication on the Teaching and Learning of English*, *26*(2), 129-154. https://doi.org/10.15639/teflinjournal.v26i2/129-154

Başman, M. (2023). A comparison of the efficacies of differential item functioning detection methods. *International Journal of Assessment Tools in Education*, *10*(1), 145–159. https://doi.org/10.21449/ijate.1135368

Berrío, Á. I., Herrera, A. N., & Gómez-Benito, J. (2019). Effect of sample size ratio and model misfit when using the difficulty parameter differences procedure to detect DIF. *The Journal of Experimental Education*, *87*(3), 367–383. https://doi.org/10.1080/00220973.2018.1435502

Çelik, M., & Özkan, Y. Ö. (2020). Analysis of differential item functioning of PISA 2015 Mathematics subtest subject to gender and statistical regions. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *11*(3), 283–301. https://doi.org/10.21031/epod.715020

Center for Educational Assessment. (2020). *Laporan hasil ujian nasional - Capaian nasional*. Pusat Penilaian Pendidikan, Kementerian Pendidikan dan Kebudayaan. https://hasilun.pusmenjar.kemdikbud.go.id/#2019!smp!capaian_nasional!99&99&999!T&T&T&T&1&!1!&

Cho, S., Suh, Y., & Lee, W. (2016). An NCME instructional module on latent DIF analysis using mixture item response models. *Educational Measurement: Issues and Practice*, *35*(1), 48–61. https://doi.org/10.1111/emip.12093

Delgado, A. R., Burin, D. I., & Prieto, G. (2018). Testing the generalized validity of the emotion knowledge test scores. *PLOS ONE*, *13*(11), e0207335. https://doi.org/10.1371/journal.pone.0207335

Desjardins, C. D., & Bulut, O. (2017). *Handbook of educational measurement and psychometrics using* R. Chapman and Hall/CRC. https://doi.org/10.1201/b20498

Effendi, E. (2011). Detecting crossing differential item functioning (CDIF): Based on item response theory. Jurnal Evaluasi Pendidikan, 2(2), 147-158. https://dx.doi.org/10.21009/JEP.022.03

Effiom, A. P. (2021). Test fairness and assessment of differential item functioning of mathematics achievement test for senior secondary students in Cross River state, Nigeria using item response theory. *Global Journal of Educational Research*, *20*(1), 55–62. https://doi.org/10.4314/gjedr.v20i1.6

French, B. F., Finch, W. H., & Immekus, J. C. (2019). Multilevel Generalized Mantel-Haenszel for differential item functioning detection. *Frontiers in Education*, *4,* 47. https://doi.org/10.3389/feduc.2019.00047

Gaberson, K. B. (1997). Measurement reliability and validity. *AORN Journal*, *66*(6), 1092–1094. https://doi.org/10.1016/S0001-2092(06)62551-9

Galli, S., Chiesi, F., & Primi, C. (2011). Measuring mathematical ability needed for "non-mathematical" majors: The construction of a scale applying IRT and differential item functioning across educational contexts. *Learning and Individual Differences*, *21*(4), 392–402. https://doi.org/10.1016/j.lindif.2011.04.005

Hadi, S., Basukiyatno, B., & Susongko, P. (2021). Differential item functioning national examination on device test mathematics high school in Central Java. *Proceedings of the 1st International Conference on Social Science, Humanities, Education and Society Development, ICONS 2020, 30 November, Tegal, Indonesia*. https://doi.org/10.4108/eai.30-11-2020.2303726

Hadi, S., Puspita, F., Ati, A. P., & Widiyarto, S. (2020). Penyuluhan dan pembelajaran karakter melalui pelaksanaan Idul Adha pada siswa SMA. *Jurnal Pemberdayaan: Publikasi Hasil Pengabdian Kepada Masyarakat*, *4*(2), 205–210. https://doi.org/10.12928/jp.v4i2.1833

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. In *Fundamentals of item response theory*. Sage Publications, Inc.

Hidajad, A. (2019). Pendidikan Indonesia: Ramai di dapur, sepi di panggung (Sebuah tinjauan perkembangan). *GETER : Jurnal Seni Drama, Tari dan Musik*, *2*(2), 1–11. https://doi.org/10.26740/geter.v2n2.p1-11

Huang, X., Wilson, M., & Wang, L. (2016). Exploring plausible causes of differential item functioning in the PISA science assessment: Language, curriculum or culture. *Educational Psychology*, *36*(2), 378–390. https://doi.org/10.1080/01443410.2014.946890

Ihsan, H. (2016). Validitas isi alat ukur penelitian konsep dan panduan penilaiannya. *PEDAGOGIA Jurnal Ilmu Pendidikan*, *13*(2), 266-273. https://doi.org/10.17509/pedagogia.v13i2.3557

James, G., James, R. C., & Davis, P. J. (1959). Mathematics dictionary. *Physics Today*, *12*(10), 50–52. https://doi.org/10.1063/1.3060526

Jusmirad, M., Angraeini, D., Faturrahman, M., Syukur, M., & Arifin, I. (2023). Implementasi literasi dan numerasi pada program MBKM dan dampaknya terhadap siswa SMP Datuk Ribandang. *Jurnal Pendidikan Indonesia*, *4*(03), 303–310. https://doi.org/10.59141/japendi.v4i03.1687

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Langer, M. M. (2008). *A reexamination of lord's wald test for differential item functioning using item response theory and modern error estimation*. Dissertation, The University of North Carolina. https://doi.org/10.17615/chn0-dz45

Leiner, J. E. M., Scherndl, T., & Ortner, T. M. (2018). How do men and women perceive a high-stakes test situation? *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02216

Ozdemir, B., & Alshamrani, A. H. (2020). Examining the fairness of language test across gender with IRT-based differential item and test functioning methods. *International Journal of Learning, Teaching and Educational Research*, *19*(6), 27–45. https://doi.org/10.26803/ijlter.19.6.2

Patricia, D. C., & Araújo, L. (2012). Differential item functioning (DIF): What functions differently for immigrant students in PISA 2009 reading items ? *JRC Publications Repository*. European Union. https://doi.org/10.2788/60811

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*(2), 197–207. https://doi.org/10.1177/014662169001400208

Retnawati, H. (2013). Pendeteksian keberfungsian butir pembeda dengan indeks volume sederhana berdasarkan teori respons butir multidimensi. *Jurnal Penelitian dan Evaluasi Pendidikan*, *17*(2), 275–286. https://doi.org/10.21831/pep.v17i2.1700

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., Gundy, C., Koller, M., Petersen, M. A., & Sprangers, M. A. G. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of Clinical Epidemiology*, *62*(3), 288–295. https://doi.org/10.1016/j.jclinepi.2008.06.003

Siegrist, M., Connor, M., & Keller, C. (2012). Trust, confidence, procedural fairness, outcome fairness, moral conviction, and the acceptance of GM field experiments. *Risk Analysis*, *32*(8), 1394–1403. https://doi.org/10.1111/j.1539-6924.2011.01739.x

Sinha, R., van den Heuvel, W. A., & Arokiasamy, P. (2013). Validity and reliability of MOS short form health survey (SF-36) for use in India. *Indian Journal of Community Medicine*, *38*(1), 22-26. https://doi.org/10.4103/0970-0218.106623

Sitepu, V. V., & Rahmawati, F. (2022). Analisis pusat pertumbuhan dan sektor ekonomi dalam mengurangi ketimpangan pendapatan. *AKUNTABEL: Jurnal Akuntansi dan Keuangan*, *19*(1), 1–12. https://download.garuda.kemdikbud.go.id/article.php?article=3275677&val=11261&title=Analisis%20pusat%20pertumbuhan%20dan%20sektor%20ekonomi%20dalam%20mengurangi%20ketimpangan%20pendapatan

Soysal, S., & Koğar, E. Y. (2021). An investigation of item position effects by means of IRT-based differential item functioning methods. *International Journal of Assessment Tools in Education*, *8*(2), 239–256. https://doi.org/10.21449/ijate.779963

Sudaryono, S. (2017). Sensitivity of differential item functioning (DIF) detection method. *Jurnal Evaluasi Pendidikan*, *3*(1), 82-94. https://doi.org/10.21009/JEP.031.07

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, *99*(1), 118–128. https://doi.org/10.1037/0033-2909.99.1.118

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In *Test validity*. (pp. 147–172). Lawrence Erlbaum Associates, Inc. https://doi.org/10.1037/14047-004

Turang, D. A. O. (2017). Pendekatan model ontologi untuk pencarian lembaga pendidikan (Studi kasus lembaga pendidikan provinsi Daerah Istimewa Yogyakarta). *Jurnal Ilmiah Teknologi Infomasi Terapan*, *3*(3), 175-182. https://doi.org/10.33197/jitter.vol3.iss3.2017.134

Uğurlu, S., & Atar, B. (2020). Performances of MIMIC and logistic regression procedures in detecting DIF. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *11*(1), 1–12. https://doi.org/10.21031/epod.531509

Ukanda, F., Othuon, L., Agak, J., & Oleche, P. (2019). Effectiveness of Mantel-Haenszel and logistic regression statistics in detecting differential item functioning under different conditions of sample size, ability distribution and test length. *American Journal of Educational Research*, *7*(11), 878–887. https://www.sciepub.com/EDUCATION/abstract/11217

Whynes, D. K., Sprigg, N., Selby, J., Berge, E., & Bath, P. M. (2013). Testing for differential item functioning within the EQ-5D. *Medical Decision Making*, *33*(2), 252–260. https://doi.org/10.1177/0272989X12465016

Yamin, M., & Syahrir, S. (2020). Pembangunan pendidikan merdeka belajar (Telaah metode pembelajaran). *Jurnal Ilmiah Mandala Education*, *6*(1), 126-136. https://doi.org/10.36312/jime.v6i1.1121

Yildirim, O. (2019). Detecting gender differences in PISA 2012 mathematics test with differential item functioning. *International Education Studies*, *12*(8), 59-71. https://doi.org/10.5539/ies.v12n8p59

Zampetakis, L. A., Bakatsaki, M., Litos, C., Kafetsios, K. G., & Moustakis, V. (2017). Gender-based differential item functioning in the application of the theory of planned behavior for the study of entrepreneurial intentions. *Frontiers in Psychology*, *8*, 451. https://doi.org/10.3389/fpsyg.2017.00451

Zukmadini, A. Y., Karyadi, B., & Rochman, S. (2021). Peningkatan kompetensi guru melalui workshop model integrasi terpadu literasi sains dan pendidikan karakter dalam pembelajaran IPA. *Publikasi Pendidikan*, *11*(2), 107-116. https://doi.org/10.26858/publikan.v11i2.18378