REID (Research and Evaluation in Education), 11(1), 2025, 59-74
Available online at: http://journal.uny.ac.id/index.php/reid

Differential item functioning analysis of Arabic language exams across gender, study specialization, and geographic region in senior high schools

Anugrah Arya Bakti*[1]; Marzuki[1]; Zulfa Safina Ibrahim[1]; Rugaya Tuanaya[1]; Nur Yusra binti Yacob[2]
[1]Universitas Negeri Yogyakarta, Indonesia
[2]Universiti Teknologi Mara Shah Alam, Malaysia
*Corresponding Author. E-mail: anugraharya.2022@student.uny.ac.id

ARTICLE INFO

Article History
Submitted:
May 30, 2025
Revised:
July 8, 2025
Accepted:
July 11, 2025

Keywords
differential item functioning (DIF); Arabic language assessment; item response theory (IRT); fairness in testing; senior high school education in Indonesia

Scan Me:

ABSTRACT

This study aims to examine the fairness of Arabic language assessment instruments used in Muhammadiyah senior high schools by detecting the presence of Differential Item Functioning (DIF) in the Final Semester Summative Test (UAS) for 12th-grade students in the Special Region of Yogyakarta during the 2023/2024 academic year. Using a descriptive quantitative design, the research analyzed student response data from 1,157 participants across 25 schools. Data collection was conducted through documentation of test blueprints, item sheets, answer keys, and student responses. Analysis was performed using the Lord and Generalized Lord methods within the framework of Item Response Theory (IRT), focusing on three demographic variables: gender, study specialization (science vs. social studies), and school region (Yogyakarta City, Sleman, Bantul, and Kulon Progo). The Rasch model was identified as the most optimal model due to its superior fit and fulfillment of key psychometric assumptions, including unidimensionality and parameter invariance. The findings indicate that several items exhibit significant DIF across all examined variables. Eleven items showed gender-based DIF, with a higher number favoring male students. Twenty-three items demonstrated DIF by study specialization, and thirty-seven items displayed DIF based on school region, with students from Yogyakarta City benefiting the most. These results suggest that the test is not fully equitable and highlight the need for item revision to ensure fairness. The study contributes theoretically to the field of educational measurement and practically to the development of fairer evaluation practices in Islamic and language education settings.

To cite this article (in APA style):
Bakti, A. A., Marzuki, M., Ibrahim, Z. S., Tuanaya, R., & binti Yacob, N. Y. (2025). Differential item functioning analysis of Arabic language exams across gender, study specialization, and geographic region in senior high schools. *REID (Research and Evaluation in Education)*, *11*(1), 59-74. https://doi.org/10.21831/reid.v11i1.85961

INTRODUCTION

Language is an essential tool for expressing thoughts, emotions, and cultural values (Alejandro, 2024; Nasution & Tambunan, 2022). In the context of Islam, Arabic holds a central position as the language of the Qur'an and the primary heritage of Islamic civilization. In Indonesia, Arabic language learning is not only taught in Islamic boarding schools (*pesantren*) but is also part of the formal curriculum, such as in Muhammadiyah schools (Sari & Hikmah, 2024). Through the ISMUBA curriculum (*A-Islam, Kemuhammadiyahan, dan Bahasa Arab*), Muhammadiyah schools aim to instill language competence while shaping students' religious character (Bakar, 2022).

The urgency of Arabic language education in Islamic schools has been empirically demonstrated. Sopian et al. (2025) found that Arabic instruction in pesantren functions not only as a language tool but also as a medium for intercultural communication and religious character

ISSN: 2460-6995 (Online)

formation in multicultural contexts. Similarly, Muttaqin et al. (2024) confirmed that students' Arabic acquisition is significantly influenced by immersive learning environments, supporting the language's crucial role in shaping Islamic identity and academic progression.

As part of the learning process, the Final Semester Examination (*Ujian Akhir Sekolah* or UAS) plays a critical role in assessing students' competence achievements (Waizah & Herwani, 2021). In the Muhammadiyah school system, Arabic is one of the core subjects in the ISMUBA curriculum. Performance in Arabic exams contributes significantly to students' final grades and can influence decisions about study specialization and graduation eligibility. Given this weight, the fairness of Arabic test items directly impacts the objectivity and equity of educational outcomes, underscoring its critical role in the assessment process (Muttaqin et al., 2024). However, attention to the fairness of test instruments is often limited. In practice, test items may contain biases that advantage or disadvantage certain groups. This carries the risk of producing inaccurate assessments and unfair educational decisions.

One approach that can be used to evaluate the fairness of a test instrument is Differential Item Functioning (DIF) analysis (Effiom, 2021; Wallin et al., 2024). This analysis aims to detect whether a test item functions differently for groups of students with equivalent ability but differing in certain characteristics such as gender, subject specialization, or geographic background. In this study, DIF analysis is conducted using Lord's Chi-Square method based on Item Response Theory (IRT). This method is used to identify whether a test item shows differential functioning between two student groups (e.g., based on gender or background), despite having equivalent ability levels. The Lord's method calculates the chi-square statistic by considering multivariate differences in item parameters, particularly focusing on discrimination and difficulty parameters, while keeping the guessing parameter constant (Downey & Stockdale, 1987). These parameter differences are then tested using a chi-square statistic with two degrees of freedom, where a significant result indicates potential bias in the item. This approach is relevant in educational settings as it provides deeper diagnostic insights into the fairness and quality of assessment instruments used in the learning process.

In the context of education in Indonesia, research explicitly analyzing the presence of DIF in Arabic language exams, especially in Muhammadiyah schools, is still limited. Yet, assessment fairness is a crucial issue to ensure that no student group is systematically disadvantaged due to item bias (Effiom, 2021; Khasawneh & Khasawneh, 2023; Tierney, 2022).

Mi'rotin and Cholil (2020) found that 32% of Arabic exam items in madrasah tsanawiyah favored one gender, directly impacting students' academic standing. These findings highlight the urgency of evaluating Arabic UAS instruments more rigorously, especially given their role in determining final grades and graduation decisions.

Previous studies have shown that differences in student performance in exams can be influenced by non-academic variables such as gender, subject specialization (science/social studies), and school region. For instance, a study by Setiawan et al. (2024) found regional bias in the National Mathematics Examination instrument. A related study by Sumin et al. (2022) also detected gender bias in the Kentucky Inventory of Mindfulness Skills (KIMS) psychology instrument.

While these studies focused on different content domains, their findings indicate that demographic variables can significantly affect test performance, particularly in the context of high-stakes assessments. In the case of Arabic Final Semester Examinations (UAS), where results contribute directly to students' academic promotion and report card grades, ensuring test fairness becomes even more critical (Wahyuni, 2022). These previously studied instruments differ in both structure and consequence from Arabic summative tests, which are closely tied to final subject grades and curricular progression, especially within Islamic education. Therefore, more targeted research on Arabic UAS is both relevant and necessary.

In this study, "subject specialization" refers to students' selected academic stream in Indonesian senior high schools under the 2013 national curriculum, either Science (*Ilmu*

Pengetahuan Alam, IPA) or Social Studies (*Ilmu Pengetahuan Sosial*, IPS). While students still study core national subjects, these streams determine additional coursework and shape their cognitive and linguistic development. Fatimah et al. (2024) observed that STEM-focused curricula in senior high schools led to more frequent engagement in analytical and disciplinary discourse than social studies programs, suggesting that such specialization can influence test performance dynamics.

Furthermore, Danuwijaya and Roebianto (2020) found that six out of 50 items in an English reading test exhibited gender-related DIF, implying that linguistic exposure and cognitive processing patterns may differ between IPA and IPS students. This supports our rationale for examining "subject specialization" in the context of Arabic UAS.

Additionally, the inclusion of regional comparisons (Yogyakarta City, Sleman, Bantul, and Kulon Progo) addresses structural differences that may arise from variations in educational resource distribution, school facilities, and curriculum supervision across districts. Although located within the same province, each district may implement educational programs differently, which could contribute to variations in student performance. Çelik and Yeşim (2020) emphasized that regional disparities can lead to differential item functioning (DIF), as evidenced in their analysis of the PISA 2015 mathematics subtest, where significant DIF was found across statistical regions in Turkey. Their findings underscore the importance of considering geographic and contextual differences in test fairness, supporting the relevance of regional DIF analysis in educational assessment.

If not properly addressed, these conditions have the potential to exacerbate learning outcome disparities and hinder the principle of fairness in education. Therefore, this study is both relevant and urgent, particularly because Arabic UAS results are widely used as benchmarks for summative evaluation and educational decision-making in Muhammadiyah senior high schools.

This research focuses on DIF analysis in Arabic UAS items used in Muhammadiyah high schools in the Special Region of Yogyakarta during the 2022/2023 academic year. The main objective is to detect whether any items systematically exhibit bias based on gender, subject specialization (science/social studies), and students' regional background. Thus, the study aims to identify potentially unfair items for future revision and improvement.

The benefits of this study span two dimensions. Theoretically, it contributes to the development of research on assessment fairness and educational measurement in language learning. Practically, the findings are expected to provide insights for teachers and educational policymakers within Muhammadiyah institutions to develop more equitable and representative evaluation instruments for all students, regardless of their demographic background.

In this study, test fairness is operationally defined through Differential Item Functioning analysis, where an item is said to function differently if it shows performance differences between two or more groups with equal ability but different characteristics (Hope et al., 2018; Liu & Rogers, 2022). This approach enables a more objective evaluation of potential bias in test items and forms an essential part of equitable assessment practices.

METHOD

This study is a descriptive quantitative research aimed at detecting the presence of Differential Item Functioning (DIF) based on gender, subject specialization (science and social studies), and school region (Yogyakarta City, Sleman, Bantul, and Kulon Progo) in the Arabic language Final Semester Summative Test (*Ujian Akhir Semester* or UAS) for 12th-grade students. This research does not involve the manipulation of any variables; rather, it utilizes existing data to identify test items that may be biased toward certain groups.

The subjects of this study consist of 12th-grade students from 25 Muhammadiyah senior high schools in the Special Region of Yogyakarta who participated in the 2023/2024 academic year's Final Semester Summative Test, totalling 1,157 respondents. Subjects were selected purposively based on the availability of student response data, without regard to initial characteristics, as DIF categories were analyzed separately according to demographic groups.

Table 1. Arabic Test Blueprint

No.	Learning Outcome	Item Indicator	Domain
1.	Able to communicate verbally about the topic "Health" *(الصحة)* with *fi'il mabni lil-majhul elements*, and the topic "Communication Media" *(وسائل االتصال)* with *adawatul istifham* elements	Presented with a question expression in a dialogue about *Ash-shihhah* (health), students are able to determine the correct response	C3 (Applying)
		Presented with a response expression in a dialogue about *Ash-shihhah*, students are able to determine the correct question	C3 (Applying)
		Presented with a dialogue about *Ash-shihhah*, students are able to translate the underlined sentence	C3 (Applying)
		Presented with a dialogue about *Ash-shihhah*, students are able to identify related information	C3 (Applying)
		Presented with an incomplete dialogue about *Ash-shihhah*, students are able to complete it using question words	C3 (Applying)
		Presented with a dialogue about *Wasailul ittishal* (communication media), students are able to complete the dialogue according to the picture	C3 (Applying)
		Presented with a question expression in a dialogue about *Wasailul ittishal*, students are able to determine the correct response	C3 (Applying)
		Presented with a dialogue about *Wasailul ittishal*, students are able to translate the underlined word	C2 (Understanding)
		Presented with a dialogue in Indonesian about *Wasailul ittishal*, students are able to translate it into Arabic	C3 (Applying)
2.	Able to read aloud, understand explicit and implicit meaning, and reflect on written texts about the topic "Health" (الصحة) with *fi'il mabni lil-majhul,* and "Communication Media" (وسائل االتصال) with *adawatul istifham*	Presented with a discourse on the topic of *Ash-shihhah*, students are able to elaborate on its content	C4 (Analyzing)
		Presented with a reading passage on *Ash-shihhah*, students are able to translate the underlined phrase	C3 (Applying)
		Presented with an image related to *Ash-shihhah*, students are able to analyze appropriate statements based on the picture	C4 (Analyzing)
		Presented with a discourse on *Wasailul ittishal*, students are able to conclude the main idea	C4 (Analyzing)
3.	Able to express ideas in writing on the topic "Health" (الصحة) with *fi'il mabni lil-majhul,* and "Communication Media" (وسائل االتصال) with *adawatul istifham*	Presented with a simple sentence about *Ash-shihhah*, students are able to identify *fi'il mabni lil-majhul*	C3 (Applying)
		Presented with an image about *Ash-shihhah*, students are able to identify related vocabulary	C3 (Applying)
		Presented with a sentence about *Ash-shihhah*, students are able to translate the underlined sentence	C3 (Applying)
		Presented with an incomplete sentence about *Ash-shihhah*, students are able to complete it with *fi'il mabni lil-majhul*	C5 (Evaluating)
		Presented with vocabulary about *Wasailul ittishal*, students are able to classify according to categories	C5 (Evaluating)
		Presented with a sentence about *Wasailul ittishal*, students are able to select a sentence with *adawatul istifham*	C4 (Analyzing)
		Presented with an image about *Wasailul ittishal*, students are able to identify the appropriate vocabulary	C3 (Applying)
		Presented with a simple sentence about *Wasailul ittishal*, students are able to identify *adawatul istifham*	C3 (Applying)
		Presented with a sentence about *Wasailul ittishal*, students are able to translate it according to the topic	C3 (Applying)
		Presented with an incomplete sentence about *Wasailul ittishal*, students are able to complete it using *adawatul istifham*	C5 (Evaluating)
		Presented with jumbled sentences about *Wasailul ittishal*, students are able to arrange them into a simple sentence	C3 (Applying)
		Presented with a picture table about *Wasailul ittishal*, students are able to classify according to categories	C5 (Evaluating)
		Presented with vocabulary about *Wasailul ittishal*, students are able to classify according to categories	C5 (Evaluating)
		Presented with an image about *Ash-shihhah*, students are able to identify the function of the tool shown	C3 (Applying)

The distribution of schools and students by region is as follows: Yogyakarta City (seven schools, 843 students), Bantul Regency (five schools, 109 students), Sleman Regency (four schools, 145 students), and Kulon Progo Regency (one school, 60 students). Gunungkidul Regency was not represented in the sample due to the unavailability of complete student response data from schools in that area at the time of data collection. In terms of demographic proportions, the sample consisted of 548 male students (47.4%) and 609 female students (52.6%). Regarding academic specialization, 649 students (56.1%) were enrolled in the science stream (IPA), and 508 students (43.9%) were in the social studies stream (IPS).

Data were collected through documentation techniques involving several materials, including the test blueprint, question sheets, answer keys, and student answer sheets. Table 1 presents the blueprint of the Arabic Final Semester Examination (UAS) used in this study, which outlines the distribution of learning outcomes, item indicators, and cognitive domains assessed. The cognitive domains are based on Bloom's Taxonomy, which classifies learning objectives into hierarchical levels ranging from lower-order to higher-order thinking skills. In this blueprint, domains such as C2 (Understanding), C3 (Applying), C4 (Analyzing), and C5 (Evaluating) are used to indicate the depth of cognitive processes expected from students.

The data were then coded and anonymized to ensure the confidentiality of student identities and to guarantee that all data were used solely for research purposes. After the data collection process, the data were analyzed quantitatively using R software (version 4.1.3).

The primary analysis in this study focuses on DIF detection using the Lord and Generalized Lord models. These techniques are employed to determine whether certain items offer different probabilities of being answered correctly by specific groups (based on gender, subject specialization, and school region), even when they possess equivalent ability levels. The results of this analysis are expected to provide objective insights into the fairness of test items for all student groups.

FINDINGS AND DISCUSSION

This study aims to detect the presence of Differential Item Functioning (DIF) in the Arabic language Final Semester Summative Test items administered at Muhammadiyah senior high schools in the Special Region of Yogyakarta during the 2023/2024 academic year, based on gender, subject specialization (science and social studies), and school region. The analysis results indicate that several test items exhibit significant differential functioning, suggesting a potential inequality in the probability of answering correctly between student groups, despite having equivalent ability levels.

Model Fit and Assumption

Table 2. Comparison of the Number of Well-Fitting Items Across Models

Category	Rasch	1PL	2PL	3PL	4PL
Not Fit	1	1	5	3	3
Fit	48	48	44	46	46

Based on Table 2, the highest number of items meeting the model fit criteria is found in the Rasch model and 1PL model, each with 48 items identified as fitting. Although both models show the same level of fit in terms of the number of fitting items, the Rasch model is selected as the most optimal model. This decision is based on its lower Akaike Information Criterion (AIC) value, which is 64,477.65 for the Rasch model compared to 65,561.38 for the 1PL model. A lower AIC value shows that the model provides the best balance between model complexity and goodness of fit to the data. Thus, the Rasch model is considered more efficient and appropriate for use in this analysis. In conclusion, the Rasch model not only excels in terms of item fit but also statistically offers the most suitable representation of the test-takers' response data.

Unidimentional

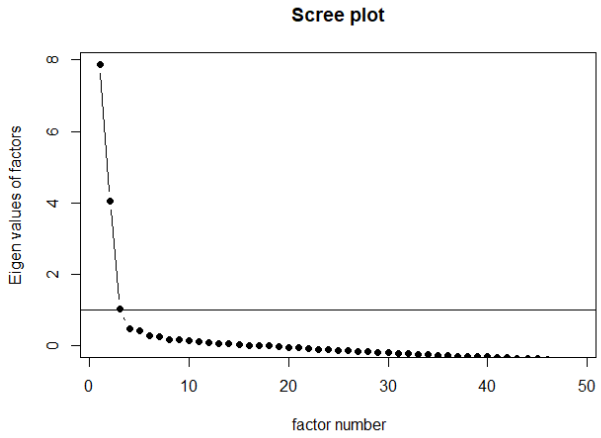

Figure 1. Scree Plot for Testing the Unidimensionality Assumption

The scree plot in Figure 1 indicates that the test meets the assumption of unidimensionality, meaning it measures only a single construct or underlying ability. This is evident from the presence of one dominant factor with a substantially higher eigenvalue of 8.666934, compared to the subsequent factors, which show a significant decline. In factor analysis, such a pattern suggests that the classity of variance in the data can be explained by a single primary dimension. If there were multiple factors with high eigenvalues, it might indicate multidimensionality; however, in this case, the displayed factor structure supports the assumption that the test measures a single dominant ability, namely, Arabic language proficiency.

Invariance of Item Difficulty Parameters

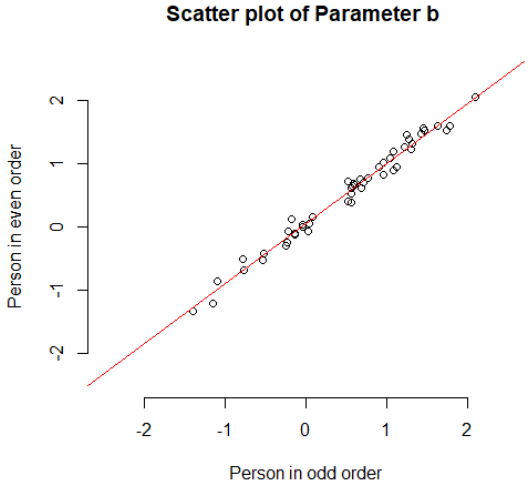

Figure 2. Scatter Plot for Testing the Invariance Assumption of Parameter b

Referring to Figure 2, the assumption of parameter invariance in measurement can be evaluated through the distribution pattern of points on the scatter plot. It is apparent that most of the points lie around the diagonal line, indicating consistency in the item difficulty parameter values across the compared groups or conditions. This pattern suggests that the differences observed are not systematic, allowing the conclusion that the invariance assumption is met. In other words, the test items demonstrate relatively stable difficulty levels between the two groups, which reinforces the validity of the instrument for fair cross-group comparisons.

Invariance of Ability Parameters

Figure 3. Scatter Plot for Testing the Invariance Assumption of Ability Parameters

Figure 3 presents a scatter plot comparing participants' ability estimates between two conditions: items in even-numbered positions and those in odd-numbered positions. The clustering of points around the diagonal line indicates that ability estimates remain relatively consistent despite changes in item sequence. This suggests that participants' ability measurement is not affected by the order of item presentation, thereby confirming that the assumption of ability parameter invariance is satisfied.

Differential Item Functioning

DIF by Gender using Lord's Method

Table 3. Gender-Based DIF Results (Lord's Method)

Test Item Number	Statistics	P-Value	Category	Test Item Number	Statistics	P-Value	Category
1	6.102	0.0135	DIF (Male)	26	0.208	0.6483	Not DIF
2	0.004	0.9525	Not DIF	27	0.1988	0.6557	Not DIF
3	0.356	0.5509	Not DIF	28	0.9152	0.3387	Not DIF
4	2.049	0.1523	Not DIF	29	0.0516	0.8203	Not DIF
5	3.902	0.0482	DIF (Female)	30	0.0035	0.9529	Not DIF
6	4.793	0.0286	DIF (Male)	31	0.8385	0.3598	Not DIF
7	1.743	0.1867	Not DIF	32	3.1249	0.0771	Not DIF
8	1.389	0.2386	Not DIF	33	2.7459	0.0975	Not DIF
9	7.304	0.0069	DIF (Female)	34	3.2342	0.0721	Not DIF
10	0.519	0.4715	Not DIF	35	2.3596	0.1245	Not DIF
11	8.071	0.0045	DIF (Male)	36	13.1504	0.0003	DIF (Female)
12	7.371	0.0066	DIF (Male)	37	1.7025	0.192	Not DIF
13	2.268	0.1321	Not DIF	38	0.7013	0.4023	Not DIF
14	2.493	0.1144	Not DIF	39	0.0534	0.8173	Not DIF
15	0.060	0.8071	Not DIF	40	3.8327	0.0503	Not DIF
16	0.442	0.5063	Not DIF	41	0.0256	0.8729	Not DIF
17	0.883	0.3473	Not DIF	42	21.8329	0	DIF (Male)
18	1.557	0.2121	Not DIF	43	0.2444	0.6211	Not DIF
19	1.559	0.2118	Not DIF	44	0.0237	0.8776	Not DIF
20	0.386	0.5347	Not DIF	45	0.0008	0.9769	Not DIF
21	3.383	0.0659	Not DIF	46	2.3603	0.1245	Not DIF
22	1.124	0.2891	Not DIF	47	3.2465	0.0716	Not DIF
23	18.227	0	DIF (Male)	48	0.0934	0.7598	Not DIF
24	19.646	0	DIF (Male)	49	0.9761	0.3232	Not DIF
25	4.767	0.029	DIF (Male)				

The results of the analysis using the Lord's method (Table 3) revealed that out of 49 test items, 11 items exhibited Differential Item Functioning (DIF) based on gender. Among these, eight items (numbers 1, 6, 11, 12, 23, 24, 25, and 42) favored male students, as they had a higher probability of answering correctly compared to female students at the same ability level. Conversely, three items (numbers 5, 9, and 36) favored female students. These findings indicate that certain items may show a tendency toward gender bias, which could stem from differences in learning styles, contextual experiences, or language elements that are more familiar to one group. The presence of DIF warrants careful consideration, as items that are intended to be neutral may systematically benefit one group over another.

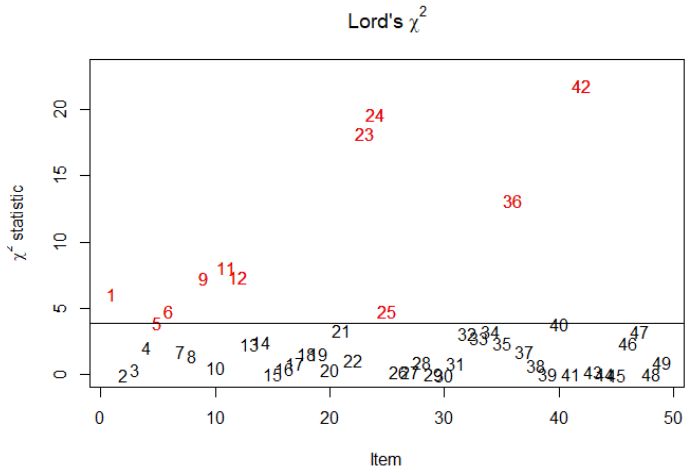

Figure 4. Plot of Gender DIF Detection Results Using Lord's Method

Figure 4 displays a visualization of the chi-square statistics from the Lord's test for each item in detecting gender-based DIF. Red dots represent items with chi-square values exceeding the significance threshold, which in this context indicates significant differences in item functioning between gender groups. It is evident that items numbered 1, 5, 6, 9, 11, 12, 23, 24, 25, 36, and 42 show relatively high chi-square statistics, positioned well above most other points, which indicate the presence of DIF in those items. This distribution pattern reinforces the finding that certain items systematically favor one gender group. Therefore, these items should be reviewed to ensure that the test instrument used is fair and does not favor a particular group.

Probability of a Student's Correct Answer Based on Gender

Figure 5 shows the Item Characteristic Curve (ICC) depicting the comparison of the probability of answering correctly between two groups, namely the reference group (males) and the focal group (females), at the same ability level (θ). The analysis results indicate that several items exhibit significant differences in their curves, which signifies the presence of Differential Item Functioning (DIF). Specifically, items 1, 6, 11, 12, 23, 24, 25, and 42 show that male students have a higher chance of answering correctly compared to female students, as the reference group's curve consistently lies above the focal group's curve. This result indicates that these items favor the male group and have the potential to introduce gender bias in the test instrument.

Conversely, items 5, 9, and 36 show the opposite pattern, where the focal group's curve is above the reference group's curve, indicating that these items favor female students. The relatively stable curve differences across the ability range for these items indicate uniform Differential Item Functioning (DIF), meaning the difference in the probability of answering correctly remains constant regardless of the ability level. Based on the analysis conducted, more items were detected to favor male students compared to female students. This finding aligns with the study by Arslan et al. (2023), which showed that male students tend to have an advantage

when taking intelligence tests. The presence of Differential Item Functioning (DIF) in these items shows that the test instrument is not yet completely free from bias, thus requiring further review of the content and structure of the items. Revision or removal of items with significant DIF should be considered to ensure the measurement of student ability is conducted fairly and equally for all groups. In contrast, items that favored female students likely involved social, relational, or familiar contextual content, which aligns with previous findings that females tend to perform better in tasks involving verbal fluency and contextual comprehension (Hope et al., 2018).

This finding aligns with the perspective of construct-irrelevant variance in test fairness theory (Messick, 1995), which posits that test bias may arise when item content interacts with test-taker characteristics in ways unrelated to the intended construct. In this case, if certain item contexts (e.g., examples related to sports, mechanical objects, or daily routines) are more familiar to one gender group, they may inadvertently give an advantage. Therefore, test developers must carefully review these DIF-flagged items to ensure that they do not systematically disadvantage one group, especially when the assessment is used for high-stakes decisions.

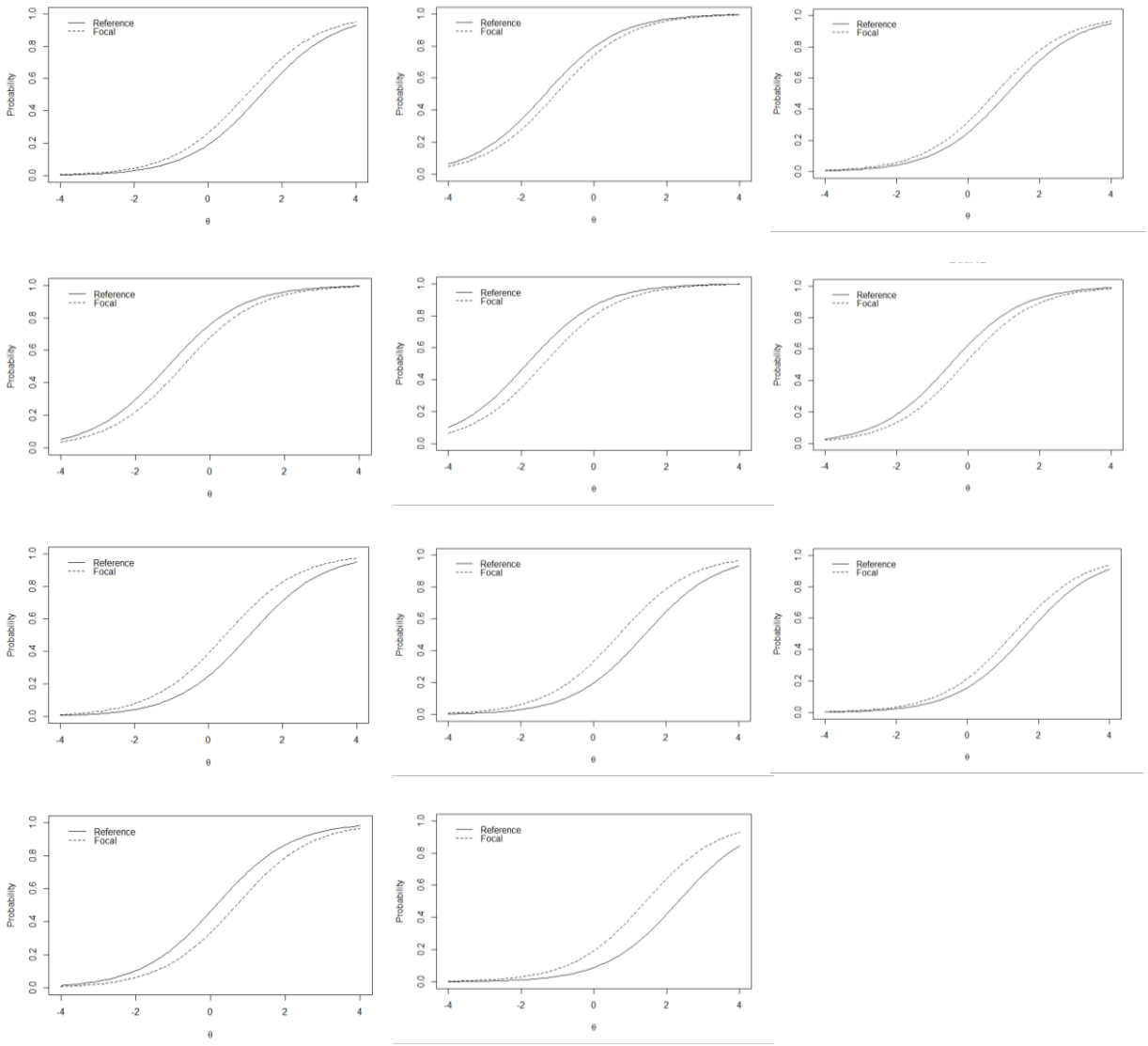

Figure 5. ICC Plot Between Males and Females for Items Detected with DIF (1, 5, 6, 9, 11, 12, 23, 24, 25, 36, and 42)

DIF by Class (Science vs. Social Studies) using Lord's method

Table 4. Class-Based DIF Results (Lord's Method)

Test Item Number	Statistics	P-Value	Category	Test Item Number	Statistics	P-Value	Category
1	0.3484	0.555	Not DIF	26	13.9423	0.0002	DIF (Science)
2	10.11	0.0015	DIF (Social)	27	2.7259	0.0987	Not DIF
3	10.1604	0.0014	DIF (Social)	28	2.1942	0.1385	Not DIF
4	13.2096	0.0003	DIF (Social)	29	4.3951	0.036	DIF (Science)
5	7.4772	0.0062	DIF (Social)	30	4.5283	0.0333	DIF (Science)
6	9.2886	0.0023	DIF (Social)	31	0.1534	0.6953	Not DIF
7	2.4334	0.1188	Not DIF	32	10.1705	0.0014	DIF (Science)
8	0.975	0.3234	Not DIF	33	2.6243	0.1052	Not DIF
9	5.2991	0.0213	DIF (Social)	34	11.6913	0.0006	DIF (Science)
10	2.535	0.1113	Not DIF	35	10.7275	0.0011	DIF (Science)
11	7.3065	0.0069	DIF (Social)	36	0.0006	0.9808	Not DIF
12	14.3284	0.0002	DIF (Social)	37	2.4205	0.1198	Not DIF
13	11.4851	0.0007	DIF (Social)	38	0.8987	0.3431	Not DIF
14	7.6625	0.0056	DIF (Social)	39	2.2895	0.1302	Not DIF
15	5.7687	0.0163	DIF (Social)	40	0.1907	0.6623	Not DIF
16	0.3898	0.5324	Not DIF	41	24.5921	0	DIF (Science)
17	0.9621	0.3267	Not DIF	42	8.832	0.003	DIF (Science)
18	1.289	0.2562	Not DIF	43	0.7913	0.3737	Not DIF
19	3.2788	0.0702	Not DIF	44	1.2988	0.2544	Not DIF
20	8.6808	0.0032	DIF (Social)	45	3.7062	0.0542	Not DIF
21	11.2176	0.0008	DIF (Social)	46	1.4897	0.2223	Not DIF
22	3.1321	0.0768	Not DIF	47	1.8433	0.1746	Not DIF
23	3.6493	0.0561	Not DIF	48	5.9902	0.0144	DIF (Science)
24	19.3727	0	DIF (Science)	49	1.3443	0.2463	Not DIF
25	1.3014	0.254	Not DIF				

The analysis of differences based on students' class revealed that 23 out of 49 items exhibited significant Differential Item Functioning (DIF) (Table 4). Among these, 13 items (numbers 2, 3, 4, 5, 6, 9, 11, 12, 13, 14, 15, 20, and 21) tended to favor students from the Social Studies (IPS) class, while 10 items (numbers 24, 26, 29, 30, 32, 34, 35, 41, 42, and 48) favored students from the Science (IPA) class. These findings indicate an imbalance in the test items with respect to the students' academic backgrounds. IPS students performed better on several items that are likely related to social or linguistic contexts closer to their field of study, whereas IPA students tended to excel on items requiring structural precision or systematic understanding typical of a scientific approach. This imbalance suggests that the distribution of item contexts and cognitive demands needs to be reconsidered to avoid bias toward certain academic backgrounds.

IPS students performed better on several items that are likely related to linguistic or contextual content, which aligns with the curriculum focus and verbal strengths commonly emphasized in social sciences. In contrast, IPA students excelled on items requiring structured reasoning, symbolic decoding, or systematic interpretation, consistent with the analytical emphasis of science curricula. This pattern echoes findings from previous studies that show that students' academic orientation can influence how they interpret and respond to test items, particularly when items reflect the discourse style or knowledge representation typical of one academic field (Effiom, 2021).

In addition, a study by Acar (2012) using the Hierarchical Generalized Linear Model (HGLM) on Science and Social Studies subtests revealed that a greater number of DIF items appeared in the Social Studies subtest compared to the Science subtest. This indicates that the context or structure of test items that aligns more closely with a particular academic specialization tends to generate bias.

Theoretically, these findings support the concept of content-structure bias as proposed by Camilli and Shepard (1994), which suggests that bias arises when the context or structure of an item aligns more closely with the academic background of certain students, thereby increasing the likelihood of DIF. Therefore, it is strongly recommended to review the items that show high DIF (items 4, 12, 24, 26, and 41), particularly by analyzing their context and cognitive demands, in order to achieve a more balanced distribution of items across study specialization.

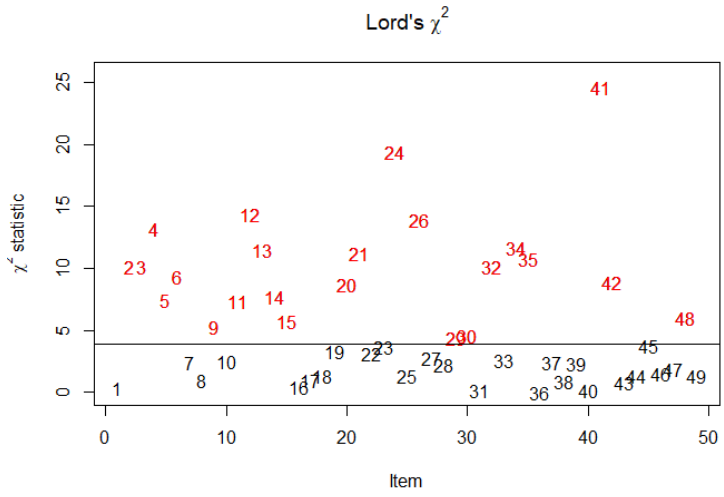

Figure 6. Plot of Class (Science vs. Social Studies) DIF Detection Results Using Lord's Method

These results are supported by Figure 6, which displays the χ^2 statistics for each test item. Items showing significant differences between the groups (based on IPA and IPS classes) are marked with red figures that stand prominently above the baseline. It can be seen that items with significant DIF (e.g., items 4, 12, 24, 26, 41) have high χ^2 values, indicating marked differences in item characteristics between IPA and IPS students, even though both groups have comparable ability levels.

In contrast, items represented by black figures near the horizontal axis show no significant difference between the groups. The distribution of DIF items is not evenly spread across the item numbers, indicating that this imbalance is not systematic by item order but rather related to the content or type of cognitive demands of each item. Therefore, revision or reorganization of the items showing indications of bias is necessary to improve the fairness of the assessment across classes.

DIF by Region using Lord's Method

The analysis using the Generalized Lord method identified that 37 out of 49 items exhibited Differential Item Functioning (DIF) based on students' school regions (Table 5). Students from Yogyakarta City had an advantage on 16 items (numbers 4, 7, 8, 9, 11, 12, 14, 15, 16, 17, 19, 23, 24, 28, 44, and 47), while students from Sleman Regency excelled on nine items (numbers 25, 26, 33, 36, 37, 38, 39, 41, and 46). Furthermore, students from Bantul Regency performed better on 8 items (numbers 2, 6, 13, 18, 21, 22, 30, and 35), and students from Kulon Progo Regency on four items (numbers 27, 29, 31, and 40). The remaining 11 items showed no significant differences between regions. These findings suggest that geographical groupings may be associated with differences in item functioning. However, Differential Item Functioning (DIF) reflects internal characteristics of students (such as prior learning experiences, test-taking strategies, or latent abilities) rather than external factors like teaching quality or local culture. Therefore, attributing regional DIF to contextual disparities requires further investigation and supporting evidence (Çelik & Yeşim, 2020; Jones, 2019; Paek, 2018).

Table 5. Region-Based DIF Results (Generalized Lord's Method)

Test Item Number	Statistics	P-Value	Category	Test Item Number	Statistics	P-Value	Category
1	2.3919	0.4951	Not DIF	26	101.4358	0	DIF
2	10.1465	0.0174	DIF	27	20.2308	0.0002	DIF
3	1.0736	0.7835	Not DIF	28	15.9742	0.0011	DIF
4	13.1628	0.0043	DIF	29	8.0705	0.0446	DIF
5	2.2169	0.5286	Not DIF	30	9.6676	0.0216	DIF
6	14.2496	0.0026	DIF	31	12.1118	0.007	DIF
7	8.8169	0.0318	DIF	32	0.2207	0.9742	Not DIF
8	13.7281	0.0033	DIF	33	8.2296	0.0415	DIF
9	24.3931	0	DIF	34	0.6272	0.8902	Not DIF
10	5.5894	0.1334	Not DIF	35	5.0248	0.17	DIF
11	28.0432	0	DIF	36	19.2958	0.0002	DIF
12	14.1529	0.0027	DIF	37	40.9164	0	DIF
13	17.2346	0.0006	DIF	38	74.6058	0	DIF
14	19.2301	0.0002	DIF	39	48.7003	0	DIF
15	36.157	0	DIF	40	68.3623	0	DIF
16	21.5833	0.0001	DIF	41	69.6576	0	DIF
17	29.415	0	DIF	42	5.3443	0.1483	Not DIF
18	9.4319	0.0241	DIF	43	3.5383	0.3158	Not DIF
19	13.7199	0.0033	DIF	44	9.1769	0.027	DIF
20	6.3026	0.0978	Not DIF	45	4.8128	0.186	Not DIF
21	25.4305	0	DIF	46	14.8102	0.002	DIF
22	15.004	0.0018	DIF	47	11.6394	0.0087	DIF
23	48.0606	0	DIF	48	21.7236	0.0001	DIF
24	17.0486	0.0007	DIF	49	0.3187	0.9565	Not DIF
25	12.7647	0.0052	DIF				

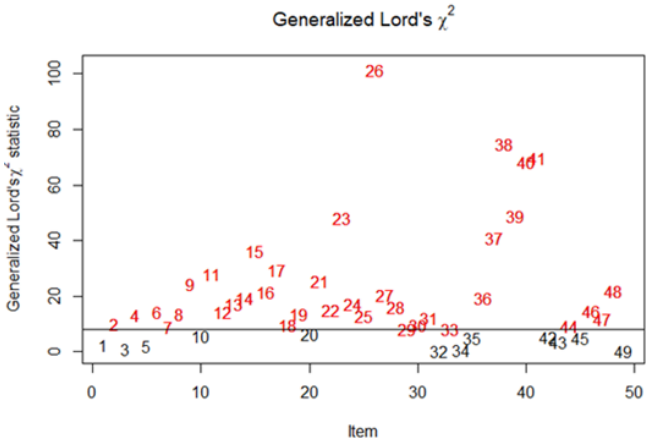

Figure 7. Plot of Region DIF Detection Results Using Generalized Lord's Method

The DIF analysis is supported by the Generalized Lord's χ^2 graph shown in Figure 7, which illustrates the magnitude of the statistic for each item in detecting DIF based on students' regions. It is evident that many items have high χ^2 values (marked with red figures well above the baseline), indicating significant differences in item characteristics between regions, even though students' abilities are comparable.

For example, item number 26 shows the highest Generalized Lord's χ^2 value, indicating the most pronounced difference in item functioning by region. Similarly, items such as 38, 37, 39, 41, and 46 display high deviations that support previous findings of relative advantages for certain student groups. Conversely, items near the horizontal axis (e.g., items 1, 3, 5, and 49) can be considered to function equivalently across all regions.

This pattern shows that imbalance is not only caused by academic background but is also influenced by contextual regional factors. This highlights the importance of designing test items that consider the diversity of students' social and geographical backgrounds to ensure the assessment instrument is truly fair and representative for all groups.

Table 6. Summary of the Number of Items Containing DIF Based on Region

Item	Total	Region
4, 7, 8, 9, 11, 12, 14, 15, 16, 17, 19, 23, 24, 28, 44, 47	16	Yogyakarta City
2, 6, 13, 18, 21, 22, 30, 35	8	Bantul
25, 26, 33, 36, 37, 38, 39, 41, 46	9	Sleman
27, 29, 31, 40	5	Kulonprogo

Based on Table 6, it can be concluded that students from Yogyakarta City gain the greatest advantage in item functioning compared to students from other regions. This is evident from the fact that 16 items favor them, the highest number among the four regions analyzed.

Meanwhile, students from Sleman Regency are advantaged in nine items, followed by students from Bantul Regency with eight items, and lastly, students from Kulon Progo Regency benefit from only five items. This uneven distribution of item advantage likely reflects disparities in educational quality, teaching practices, access to learning resources, and possibly linguistic or contextual familiarity with the test items.

The predominance of Yogyakarta City in this analysis may be explained by its relatively strong educational infrastructure and access to qualified educators, which have long been associated with better student achievement in national assessments. This is supported by (Setiawan et al., 2024), who found significant DIF favoring students from urban regions such as Yogyakarta over more rural areas like South Kalimantan in national exam items, citing contextual alignment and resource availability as potential causes. Their study further indicated that all DIF-flagged items favored the Yogyakarta region, raising questions about fairness in test content and design (Setiawan et al., 2024).

These results align with broader findings from Azzizah (2015), who documented that urban schools in Indonesia, particularly in Java, tend to outperform rural counterparts due to better infrastructure, teaching quality, and resource access. Therefore, the observed dominance of Yogyakarta City in item functioning likely reflects deeper systemic inequities that can affect student outcomes. This highlights the importance of incorporating regional equity considerations into assessment design. Specifically, reviewing item content for cultural and contextual bias is critical to ensure test fairness, especially for regional or national-scale assessments where diverse populations are evaluated using the same instrument.

CONCLUSION

This study revealed that several items in the Arabic Final Semester Summative Test administered at Muhammadiyah Senior High Schools in Yogyakarta exhibited Differential Item Functioning (DIF) across gender, academic specialization, and school region. Using the Rasch model and Lord's Chi-Square method, the analysis identified 11 items with gender-based DIF, 23 items with DIF based on subject specialization, and 37 items with region-based DIF. These findings indicate that some test items did not function equivalently for students of comparable abilities but from different demographic groups, potentially affecting fairness. Gender-based DIF reflects construct-irrelevant variance (Messick, 1995), where item content may align differently with cognitive traits across male and female students (Arslan et al., 2023; Hope et al., 2018). DIF based on academic specialization further supports content-structure bias theory (Camilli & Shepard, 1994), highlighting the influence of students' learning tracks on item accessibility (Acar, 2012; Effiom, 2021). Notably, regional DIF was the most prevalent, suggesting contextual disparities in school environments may influence student responses, consistent with findings by

Setiawan et al. (2024) and Azzizah (2015). These results reinforce the need for inclusive and equitable test design that accounts for diverse student backgrounds to maintain the validity and fairness of high-stakes assessments.

ACKNOWLEDGMENT

We sincerely thank the Basic and Secondary Education Council and Non-Formal Education Leadership of the Muhammadiyah Regional Leadership of the Special Region of Yogyakarta (DIKDASMEN PNF PWM DIY) for providing the test instruments and data for this study. I also appreciate the students who participated and contributed their answer sheets, as well as everyone who supported and assisted me throughout the research process.

DISCLOSURE STATEMENT

The authors do not have any potential conflicts of interest to disclose.

FUNDING STATEMENT

This work does not receive funding.

ETHICS APPROVAL

There is no ethics approval needed because this study used secondary data obtained through documentation of existing test instruments and student answer sheets. No direct intervention, experimental treatment, or involvement of human participants was conducted during the research process.

REFERENCES

Acar, T. (2012). Determination of a differential item functioning procedure using the hierarchical generalized linear model. *Sage Open*, *2*(1), 1-8. https://doi.org/10.1177/2158244012436760

Alejandro, J. (2024). The role of language in thought formation and personality. *International Journal of Multidisciplinary Sciences*, *2*(4), 356–367. https://doi.org/10.37329/ijms.v2i4.3759

Arslan, D., Tamul, Ö. F., Şahin, M. D., & Sak, U. (2023). Effects of gender norms on intelligence tests: Evidence from ASIS. *Journal of Pedagogical Research*, *7*(5), 374–384. https://doi.org/10.33902/JPR.202323599

Azzizah, Y. (2015). Socio-economic factors on Indonesia education disparity. *International Education Studies*, *8*(12), 218-230. https://doi.org/10.5539/ies.v8n12p218

Bakar, H. I. A. (2022). Implementation of Islamic values in ISMUBA curriculum to form a Rabbani generation at Muhammadiyah Sidareja High School. *Journal of Islamic Education and Innovation*, *3*(2), 78–85. https://doi.org/10.26555/jiei.v3i2.6616

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

Çelik, M., & Yeşim, Ö. 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

Danuwijaya, A. A., & Roebianto, A. (2020). Performance differences by gender in English reading test. *Jurnal Penelitian dan Evaluasi Pendidikan*, *24*(2) 190-197. https://doi.org/10.21831/pep.v24i2.34344

Downey, R. G., & Stockdale, M. S. (1987). Computer programs to compute lord's item bias statistic for a three-parameter ICC. *Educational and Psychological Measurement*, *47*(3), 637–641. https://doi.org/10.1177/001316448704700313

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

Fatimah, S., Rusilowati, A., Cahyono, E., & Rokhmaniyah. (2024). STEM learning in higher education: A comparative study of science curriculum in Singapore and Indonesia. *International Journal of Scientific Multidisciplinary Research*, *2*(8), 1003–1030. https://doi.org/10.55927/ijsmr.v2i8.11048

Hope, D., Adamson, K., McManus, I. C., Chis, L., & Elder, A. (2018). Using differential item functioning to evaluate potential bias in a high stakes postgraduate knowledge based assessment. *BMC Medical Education*, *18*(1), 64. https://doi.org/10.1186/s12909-018-1143-0

Jones, R. N. (2019). Differential item functioning and its relevance to epidemiology. *Current Epidemiology Reports*, *6*(2), 174–183. https://doi.org/10.1007/s40471-019-00194-5

Khasawneh, M. A. S., & Khasawneh, Y. J. A. (2023). *Achieving assessment equity and fairness: Identifying and eliminating bias in assessment tools and practices*. Preprints, 2023060730. https://doi.org/10.20944/preprints202306.0730.v1

Liu, X., & Rogers, H. J. (2022). Treatments of differential item functioning: A comparison of four methods. *Educational and Psychological Measurement*, *82*(2), 225–253. https://doi.org/10.1177/00131644211012050

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Mi'rotin, S., & Cholil, M. (2020). Analisis bias gender pada soal ujian Bahasa Arab di madrasah tsanawiyah. *An Nabighoh: Jurnal Pendidikan dan Pembelajaran Bahasa Arab*, *22*(02), 191-210. https://doi.org/10.32332/an-nabighoh.v22i02.2232

Muttaqin, I., Bakheit, B. M., & Hasanah, M. (2024). Arabic language environment for Islamic boarding school student language acquisition: Capturing language input, interaction, and output. *Al-Hayat: Journal of Islamic Education*, *8*(3), 891–907. https://doi.org/10.35723/ajie.v8i3.624

Nasution, F., & Tambunan, E. E. (2022). Language and communication. *International Journal of Community Service (IJCS)*, *1*(1), 01–10. https://doi.org/10.55299/ijcs.v1i1.86

Paek, I. (2018). Understanding differential item functioning and item bias in psychological instruments. *Psychology and Psychotherapy: Research Study*, *1*(3). https://doi.org/10.31031/PPRS.2018.01.000514

Sari, R. R., & Hikmah, K. (2024). Implementation of Arabic language learning activities at the Muhammadiyah 2 Sidoarjo High School Boarding School. *Al Mi'yar: Jurnal Ilmiah Pembelajaran Bahasa Arab dan Kebahasaaraban, 7*(2), 1-9. https://doi.org/10.21070/ups.5350

Setiawan, A., Kassymova, G. K., Mbazumutima, V., & Agustyani, A. R. D. (2024). Differential item functioning of the region-based national examination equipment. *REID (Research and Evaluation in Education)*, *10*(1), 99–113. https://doi.org/10.21831/reid.v10i1.73270

Sopian, A., Abdurahman, M., Ali Tantowi, Y., Nur Aeni, A., & Maulani, H. (2025). Arabic language learning in a multicultural context at pesantren. *Jurnal Pendidikan Islam*, *11*(1), 77–89. https://doi.org/10.15575/jpi.v11i1.44104

Sumin, S., Sukmawati, F., & Nurdin, N. (2022). Gender differential item functioning on the Kentucky Inventory of Mindfulness Skills instrument using logistic regression. *REID (Research and Evaluation in Education)*, *8*(1), 55–66. https://doi.org/10.21831/reid.v8i1.50809

Tierney, R. D. (2022). *Fairness in educational testing and assessment*. Routledge. https://doi.org/10.4324/9781138609877-REE35-1

Wahyuni, A. (2022). Detection of gender biased using DIF (Differential Item Functioning) analysis on item test of school examination Yogyakarta. *Jurnal Evaluasi Pendidikan*, *13*(1), 46–49. https://doi.org/10.21009/jep.v13i1.26554

Waizah, N., & Herwani, H. (2021). Penilaian pengetahuan tertulis dalam kurikulum 2013. *Tafkir: Interdisciplinary Journal of Islamic Education*, *2*(2), 207–228. https://doi.org/10.31538/tijie.v2i2.54

Wallin, G., Chen, Y., & Moustaki, I. (2024). DIF analysis with unknown groups and anchor items. *Psychometrika*, *89*(1), 267–295. https://doi.org/10.1007/s11336-024-09948-7