JURNAL EDUSCIENCE (JES)
p-ISSN : 2303 - 355X | e-ISSN : 2685 - 2217
PUBLISHED BY : LPPM of UNIVERSITAS LABUHANBATU

Analysis of Differential Item Functioning (DIF) in the Final School Assessment Instrument for the Chemistry Subject

Sangkala Maros[1], Mansyur[2], Iwan Suhardi[3]

[1,2,3] Department of Research and Evaluation Education, Universitas Negeri Makassar, Sulawesi Selatan, Indonesia
*Email: sangkalausa@gmail.com

ARTICLE INFO

Keywords:
Differential Item Functioning
Item Response Theory
Assessment Fairness
Chemistry Instrument

ABSTRACT

Purpose – Fair assessment is a fundamental pillar for ensuring that evaluation results accurately reflect students' abilities without bias. This study aims to identify Differential Item Functioning (DIF) in the Final School Assessment (PAS) instrument for the Chemistry subject, based on four demographic variables: gender, family economic status, residential location, and school of origin.

Methodology – This study uses a quantitative design with a descriptive-exploratory approach. The research subjects were responses from 1,840 twelfth-grade students at high schools in Maros Regency. Data analysis was conducted using an Item Response Theory (IRT) approach in R (version 2024.4.2, Build 764). After assumption tests (unidimensionality and local independence) and a model fit test, the 1-Parameter Logistic (1PL) Model was selected as the most suitable. DIF detection was performed using Raju's Area Measures.

Findings – The analysis results showed that the assumptions of unidimensionality and local independence were met. Out of 30 items, five showed statistically significant DIF ($p < 0.05$): one item based on gender (Item 11), one based on residential location (Item 26), and three based on economic status (Items 19, 23, and 24). No items showed DIF by school of origin. Although statistically detected, the effect size analysis showed that all DIF items fell into Category 'A' (negligible) according to ETS criteria.

Contribution – This study provides empirical evidence regarding the fairness of an assessment instrument developed by a teacher association (MGMP). It highlights how non-academic factors can manifest as measurable differences in performance. This study affirms the importance of DIF analysis as a standard procedure in the quality assurance of assessment instruments to maintain fairness for all students.

Received 17 July 2025; Received in revised form 8 September 2025; Accepted 10 Maret 2026
Jurnal Eduscience (JES) Volume 13 No. 2 (2026)
Available online 30 April 2026

INTRODUCTION

Assessment is an integral element in the learning cycle, serving as a crucial mechanism for measuring student achievement. For results to be credible, the process must be based on the principle of fairness. A fair assessment means the evaluation process is designed to be free of bias, so it does not provide a systematic advantage or disadvantage to students based on irrelevant individual characteristics, such as gender, socioeconomic background, or culture. When a test is unfair, the validity of inferences drawn from its results is compromised, potentially leading to discrimination, and the source of unfairness is often the instrument's quality (Istiqlal et al., 2024; Safitri, Lestarani, et al., 2024; Safitri & Ansyari, 2024).

In modern psychometric terminology, a fair instrument is free from bias at the item level, a condition analyzed through the detection of Differential Item Functioning (DIF). Conceptually, DIF occurs when test-takers from different groups (e.g., males and females) with the same latent ability show different item-response probabilities. Within the framework of Item Response Theory (IRT), this is operationally defined as a condition in which the Item Characteristic Curve (ICC) for an item differs across groups. The presence of DIF indicates that test scores may be contaminated by additional attributes beyond the primary ability being measured, known as "nuisance dimensions." It is important to emphasize that DIF is a statistical phenomenon detectable through data analysis and is not automatically synonymous with bias; according to (Rust et al., 2020), detecting DIF is the first step in identifying a potential problem, whereas bias itself refers to the unfairness caused by irrelevant factors. The causes of DIF can originate in internal item factors (such as format, diction, or visual elements) and in external factors related to test-takers' characteristics.

The urgency to detect DIF becomes highly relevant in the context of the Final School Chemistry Assessment instrument in Maros Regency for the 2023/2024 academic year. The instrument, developed by the Subject Teacher Conference (MGMP), consists of 30 items administered online and has never been analyzed for potential DIF. The risk of bias in this Chemistry test is considered high due to its abstract concepts and the use of examples that may not be equally familiar to all students, a concern reinforced by previous research by Rizky (2022) and Nursaida (2018), who successfully found gender bias in other chemistry instruments. Therefore, allowing DIF to go undetected would undermine the integrity of the assessment system, produce misleading information about student abilities, and potentially harm specific student groups, making this analysis an essential step to ensure a fair and accurate evaluation.

The final school assessment for Chemistry in Maros Regency during the 2023/2024 academic year uses an instrument developed by the Subject Teacher Conference (MGMP), which plays a significant role in school-level evaluation. However, herein lies a significant research gap: no empirical study has specifically analyzed Differential Item Functioning (DIF) in an instrument designed and used at this local level. This situation becomes crucial because, unlike national-level instruments that undergo a multi-layered psychometric validation process, MGMP instruments are often directly implemented without in-depth trials.

In fact, the potential for bias in this local instrument is quite high, given the abstract nature of chemistry and the uneven familiarity with question contexts among students from diverse demographic backgrounds in Maros Regency. Without DIF analysis, the validity and fairness of this assessment cannot be ascertained, which can have direct impacts on student outcomes, teacher performance evaluations, and the mapping of educational quality at the regional level. Therefore, this research not only fills the gap in the study of local instruments but also underscores the academic urgency of ensuring that the principle of assessment fairness is truly applied at the school level, where assessment results have real consequences.

Specifically, this research aims to identify the presence of Differential Item Functioning (DIF) in the Final School Assessment (PAS) for Chemistry instrument developed by the MGMP of Maros Regency, through a quantitative analysis to test whether the test items function unequally based on four primary demographic variables: gender, economic status, place of residence, and school of origin. Beyond mere statistical identification, this research makes a significant contribution by addressing a research gap in the psychometric quality of instruments developed at the local level (MGMP), which often escape in-depth analysis. Theoretically, these findings enrich the literature on educational measurement and psychometrics by providing empirical data on assessment fairness within the specific context of Indonesia. In practice, this research provides valuable, evidence-based feedback for teachers, item developers in the Maros Regency

MGMP, and policymakers. The results can be directly used to evaluate and revise test items, thereby ensuring a fairer and more accurate assessment process for all students.

METHODOLOGY

Research Design

A quantitative approach serves as the primary foundation, as this study focuses on the statistical analysis of numerical data to detect indicators of Differential Item Functioning (DIF). This analysis is based on the framework of Item Response Theory (IRT), a modern psychometric approach that allows for an in-depth examination of item properties. Specifically, after conducting a model fit test, this research uses the one-parameter logistic (1PL) model as the basis for the analysis. The DIF detection process uses Raju's Area Measures method, which statistically estimates the area difference between the Item Characteristic Curves (ICCs) for the reference and focal groups.

Participant

The population in this study consisted of all 12th-grade high school students in Maros Regency who participated in the final school assessment for chemistry during the 2023/2024 academic year. The sampling technique used was total sampling, collecting all valid responses from the population. From this process, a final sample size of 1,840 student responses was obtained, with the following demographic distribution: 917 male and 921 female students; 964 students from urban areas and 876 from rural areas; 928 students from affluent families and 912 from less affluent families; and 929 students from public junior high schools and 911 from private junior high schools.

Data Collection

Data collection in this study was conducted using the documentation method, in accordance with applicable research ethics procedures. The process began with the submission of official permits to the relevant education agency and the principals of each participating institution, after which approval was granted for limited access to the relevant data sources. To maintain confidentiality, all collected data was anonymized by removing personal student identifiers and replacing them with unique codes. Demographic data was obtained through access to the National Education Data Pokok (DAPODIK) system, while response data was gathered from the schools' internal databases. Subsequently, a data verification mechanism was implemented to ensure accuracy and completeness, including cross-checking and data cleaning to handle invalid or missing responses before the data were ready for analysis.

Instrument

The analyzed final school assessment instrument for the subject of chemistry is the result of a consensus from the Subject Teacher Conference (MGMP) for Chemistry in Maros Regency for the 2023/2024 academic year. Its development process followed structured stages to ensure quality, validity, and reliability, involving several teams with specific tasks, including a Steering Committee, a Grid Development Team, an Item Writing Team, and a Validator and Reviewer Team. Analysis of this instrument reveals a highly structured and high-quality design, examined through three main pillars.

First, the instrument has a comprehensive and proportional content coverage to ensure content validity, where out of a total of 30 items, the distribution is 8 items (26.7%) for Basic Chemistry & Stoichiometry, 8 items (26.7%) for Physical Chemistry (Rate, Equilibrium, Acid-Base), 4 items (13.3%) for Electrochemistry, 5 items (16.7%) for Organic Chemistry & Polymers, and 5 items (16.7%) for Elemental Chemistry, Colloids & Biomolecules. Second, its cognitive level distribution shows an excellent balance with an orientation towards Bloom's Taxonomy, dominated by Higher-Order Thinking Skills (HOTS). The distribution consists of 30% Lower-Order Thinking Skills (LOTS), comprising 3 items for Remembering (C1) and 6 items for Understanding (C2), and 70% HOTS, comprising 11 items for Applying (C3) and 10 items for Analyzing (C4). Third, the quality of the item indicators is designed with precision, using measurable operational verbs

and concrete stimuli, which directly enhances the instrument's construct validity. Overall, the combination of representative content coverage, an emphasis on HOTS, and precise indicators makes this instrument's blueprint ideal and strongly supported by modern assessment theory.

Data Analysis

For the Differential Item Functioning (DIF) analysis, the research sample, comprising 1,840 student responses, was dichotomized based on four demographic variables to ensure transparent comparisons across groups. The grouping criteria used were: Gender, divided into Male and Female groups; Place of Residence, divided by geographical location, namely Urban and Rural; Economic Status, divided based on parental income, where the "Affluent" group has an income above IDR 2,000,000 per month and the "Less Affluent" group has an income below IDR 2,000,000 per month; and Junior High School Origin, divided by the school's status, namely Public and Private. Based on these criteria, the distribution of respondents per group.

Table 1. Demographic Characteristics of the Sample

Demographic Variable	Category	Total (N=1,840)	Percentage (%)
Gender	Male	918	49.8%
	Female	922	50.2%
Place of Residence	Urban	964	52.4%
	Rural	876	47.6%
Economic Status	Affluent	928	50.4%
	Less Affluent	912	49.6%
Junior High School Origin	Public	929	50.5%
	Private	911	49.5%

Data analysis was performed using an Item Response Theory (IRT) approach in R (version 2024.4.2, Build 764). The analysis stages were as follows:

IRT Assumption Test

The assumptions of unidimensionality and local independence were tested using Confirmatory DETECT analysis with the SIRT package. The criterion used was a DETECT index value < 0.20, indicating essential unidimensionality.

Model Fit Test

A comparison was made between the 1-Parameter Logistic (1PL), 2-Parameter (2PL), and 3-Parameter Logistic (3PL) models. The 3PL model proved unstable (the Hessian matrix was not positive definite) and was not used further. The choice between the 1PL and 2PL models was based on AIC, BIC, and the Likelihood Ratio Test (LRT). Based on this analysis, the 1PL model was selected as the most robust and stable model. The decision to choose the 1PL model over the 2PL model is based on strong academic considerations and statistical evidence. This choice is grounded in the principle of parsimony (Occam's Razor), which states that a simpler model is preferred if a more complex model does not provide a statistically significant improvement in fit—based on the Likelihood Ratio Test (LRT) comparing the two models, the p-value was 0.364. This value exceeds the significance level ($\alpha = 0.05$), indicating that adding the discrimination parameter to the 2PL model does not yield a statistically significant improvement in fit relative to the 1PL model. This decision is also supported by a comparison of the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), where the 1PL model shows lower values (AIC=60057.98, BIC=60223.50) compared to the 2PL model (AIC=60085.91, BIC=60416.96). Lower AIC/BIC values indicate that the 1PL model has the best balance between fit and complexity and is expected to be more stable and have better generalization performance for the current sample size (Guo et al., 2022; Gyamfi & Acquaye, 2023). Thus, the 1PL model was chosen as it is the most efficient and statistically adequate for analyzing this instrument's data.

DIF Detection

Once the 1PL model was established, DIF analysis was conducted for each item using Raju's Area Measures, implemented via the difR package. An item was considered to show DIF if the p-value was less than 0.05.

Effect Size Analysis

For items identified with DIF, their practical significance was analyzed using the Educational Testing Service (ETS) classification scheme: Category 'A' (negligible), 'B' (slight to moderate), and 'C' (moderate to large).

FINDINGS

Assumption and Model Fit Tests

To ensure the suitability of the measurement model, a Confirmatory DETECT analysis and model fit evaluation were conducted. These procedures are essential in item response theory (IRT) to assess whether the data meet key assumptions such as unidimensionality and local independence, as well as to determine the most appropriate model for further analysis. The results of these analyses are summarized in Table 2.

Table 2. Summary of Confirmatory DETECT and Model Fit Analysis

Analysis Type	Indicator	Value	Interpretation
Confirmatory DETECT	DETECT	-0.206	Supports unidimensionality
	ASSI	-0.310	Indicates local independence
	RATIO	-0.365	Confirms weak multidimensional structure
Model Fit Comparison	AIC (1PL)	60057.98	Lower (better fit)
	BIC (1PL)	60223.50	Lower (better fit)
	AIC/BIC (2PL)	Higher	Less optimal compared to 1PL
Likelihood Ratio Test (LRT)	p-value	0.364	2PL not significantly better than 1PL
Final Model Selection	Selected Model	1PL	Most parsimonious and appropriate

As shown in Table 2, the Confirmatory DETECT analysis provides strong evidence supporting the unidimensional structure of the instrument. The DETECT value of -0.206 is well below the commonly accepted threshold of 0.20, indicating that the test measures a single dominant latent trait. Additionally, the negative ASSI (-0.310) and RATIO (-0.365) values further reinforce the assumption of local independence among items, suggesting that item responses are not influenced by dependencies beyond the underlying construct being measured.

From a model fit perspective, the comparison between the 1-Parameter Logistic (1PL) and 2-Parameter Logistic (2PL) models reveals that the 1PL model provides a more optimal fit. This is evidenced by the lower Akaike Information Criterion (AIC = 60057.98) and Bayesian Information Criterion (BIC = 60223.50) values for the 1PL model, indicating better parsimony and efficiency in explaining the data.

Furthermore, the Likelihood Ratio Test (LRT) result shows a *p*-value of 0.364, which is not statistically significant. This indicates that the additional complexity introduced by the 2PL model does not lead to a significantly improved fit compared to the simpler 1PL model. Therefore, the 1PL model is preferred based on both statistical and theoretical considerations.

Overall, these findings confirm that the instrument satisfies key IRT assumptions and that the 1PL model is the most appropriate choice for subsequent analyses. The combination of strong unidimensionality evidence and optimal model fit supports the robustness, validity, and interpretability of the measurement model used in this study.

Differential Item Functioning (DIF) Analysis

The analysis of Differential Item Functioning (DIF) in this study was conducted using Raju's method to identify potential item bias across different demographic groups. This approach enables the examination of whether specific test items function differently for distinct groups of respondents who possess comparable underlying abilities. Therefore, the DIF analysis plays a crucial role in ensuring the fairness and validity of the measurement instrument employed in this research.

Table 3. Summary of DIF Analysis Results

Demographic Variable	Item with Indicated DIF	Z Raju Statistic	p-value	Effect Category (ETS)
Gender	Item 11	2,843	0.0045	A (Negligible)
Place of Residence	Item 26	2,218	0.027	A (Negligible)
Economic Status	Item 19	2,087	0.037	A (Negligible)
	Item 23	2,567	0.01	A (Negligible)
	Item 24	2,166	0.03	A (Negligible)
School Origin	None	-	-	-

Based on the results presented in Table 3, a total of five items were identified as exhibiting statistically significant DIF. For the gender variable, Item 11 showed a Z Raju statistic of 2.843 with a *p*-value of 0.0045, indicating a significant difference in item functioning between male and female respondents. Regarding place of residence, Item 26 was found to have significant DIF, with a Z value of 2.218 and a *p*-value of 0.027. Furthermore, for the economic status variable, three items—Item 19, Item 23, and Item 24—demonstrated significant DIF, with Z values of 2.087 ($p = 0.037$), 2.567 ($p = 0.01$), and 2.166 ($p = 0.03$), respectively.

Despite these statistically significant findings, all identified items fall within Category A according to the ETS classification, indicating that the magnitude of DIF is negligible. This suggests that although there are detectable differences in item functioning across groups, the practical impact of these differences is minimal and does not substantially threaten the fairness of the instrument. Additionally, no items were found to exhibit DIF based on school origin, implying that the instrument functions consistently across respondents from different educational backgrounds.

Overall, these findings indicate that the instrument demonstrates a high level of fairness and is largely free from substantial bias across the examined demographic variables. Nevertheless, the presence of statistically significant DIF, even at a negligible level, should be taken into consideration in future instrument development to further enhance measurement equity across diverse populations.

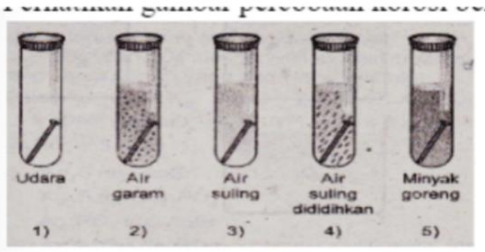
DISCUSSION

Although all detected DIF had a negligible effect size (Category A), a qualitative analysis of the flagged items provides important insights into how non-academic factors can influence student performance.

Item 11: DIF Based on Gender

A substantive analysis of Item 11 indicates a statistically significant difference in item functioning between male and female students, with the question being more difficult for male students. However, before concluding the instrument's unfairness, it is important to understand the practical significance of this finding. Although statistically significant (Z Raju = 2.843; $p < 0.05$), its effect size falls into Category 'A' (negligible) according to ETS criteria. This confirms that while a performance difference was detected, its impact on the overall fairness of the Final School Assessment (PAS) for Chemistry instrument is relatively small and does not threaten the validity of the scores.

Perhatikan gambar percobaan korosi berikut:



Paku yang mengalami korosi paling cepat terjadi pada gambar nomor …

a. 2
b. 3
c. 4
d. 5
e. 1

Figure 1. Indicated to Have DIF Based on Gender

Item 11 presents a diagram of five corrosion experiments on nails and asks students to identify the condition where corrosion occurs fastest. The correct answer (salt water) requires understanding that corrosion needs water and oxygen and is accelerated by an electrolyte. The finding that this question was easier for females can be explained by several theories: cognitive difference theory, which posits that females tend to be more thorough and detail-oriented, a cognitive style well-suited for analyzing five different visual scenarios and identifying all relevant factors. Gender Role and Socialization Theory, females are often socialized to be more compliant with rules and procedures, which, in the context of a visual practical question, encourages a systematic approach and minimizes errors. Fine Motor Skills: While not directly tested, prior lab experience requiring precision may have benefited female students, who, according to developmental psychology, tend to have superior fine motor skills. Despite the statistical advantage, the item is considered content-valid as it effectively measures the understanding of corrosion concepts.

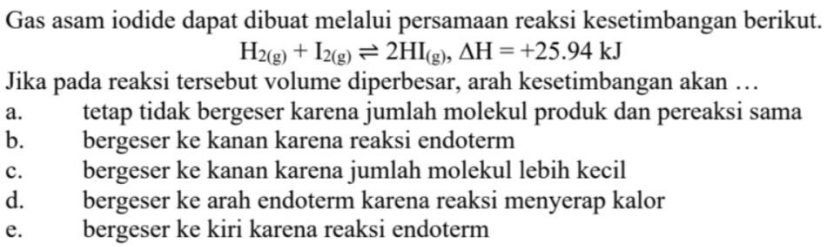
The interpretation of this finding suggests that its cause is not content bias but likely a difference in cognitive approaches and learning styles. Item 11, which requires careful observation of a corrosion experiment diagram, advantages students with a methodical approach and high perceptual speed for visual details—skills that are, on average, more common in females. In the local learning context of Maros Regency, if classroom practices do not uniformly emphasize the analysis of visual case studies or the interpretation of experimental data, students who are not naturally inclined toward that learning style (in this case, more male students) may be less trained. This difference reflects the diversity of learning styles within the student population in the area rather than an unfairness in the instrument itself.

From this finding, the practical implication for teachers and item developers at the Maros Regency MGMP is not to revise or eliminate Item 11, as the question is valid in measuring analytical skills. Instead, this result suggests the need for improvements in teaching practices. Teachers are encouraged to integrate more learning methods that accommodate various cognitive styles, for instance, by explicitly training all students in visual data interpretation and systematic observation. More broadly, this finding underscores the importance of teacher training in understanding potential differences in learning styles and how they might manifest as statistical DIF in assessments, even when the item itself is fair. Thus, the instrument's overall quality is maintained. At the same time, the focus of improvement shifts to the pedagogical domain to ensure that all students have an equal opportunity to develop the skills being tested (Lerner et al., 2021; Thiyagarajan et al., 2022).

Item 26: DIF Based on Residential Location

A substantive analysis of Item 26 reveals a statistically significant difference in performance between urban and rural students, with the question being more difficult for rural students. However, it is important to note that the practical significance of this finding is very low. Although DIF was statistically detected (Z Raju = 2.218; $p < 0.05$), its effect size is classified as Category 'A' (negligible). This indicates that even though

a performance difference was detected, this item's impact on the overall fairness of the Final School Assessment (PAS) for Chemistry instrument is considered very small and not substantial.


Gas asam iodide dapat dibuat melalui persamaan reaksi kesetimbangan berikut.

$H_{2(g)} + I_{2(g)} \rightleftharpoons 2HI_{(g)}$, $\Delta H = +25.94$ kJ

Jika pada reaksi tersebut volume diperbesar, arah kesetimbangan akan …

a. tetap tidak bergeser karena jumlah molekul produk dan pereaksi sama
b. bergeser ke kanan karena reaksi endoterm
c. bergeser ke kanan karena jumlah molekul lebih kecil
d. bergeser ke arah endoterm karena reaksi menyerap kalor
e. bergeser ke kiri karena reaksi endoterm

Figure 2. Indicated to Have DIF Based on Residential Location

Item 26 tests the understanding of Le Châtelier's Principle for the equilibrium system

$$H_2(g) + I_2(g) \rightleftharpoons 2HI(g) \quad (1)$$

when the volume is increased. The correct answer is that the equilibrium does not shift because the number of gas moles is equal on both sides of the reaction (2 moles). This item is a classic "trap" designed to distinguish conceptual understanding from mere rote memorization of the rule "if volume increases, shift to more moles".

The advantage for urban students can be explained by the gap in access to quality educational ecosystems: Urban students are more likely to be taught by teachers trained to foster conceptual reasoning rather than just memorization, aligning with constructivist learning theory. Access to learning resources, they have greater access to a variety of practice questions, including "trap" questions like this one, which hones their analytical skills and ability to filter relevant information (such as ignoring the provided ΔH data). This gap is not about intellectual capacity but an accumulation of advantages from better educational capital. The item is considered high-quality for testing critical thinking.

The interpretation of this finding points to a gap in Opportunity to Learn (OTL), rather than content bias in the question. Item 26 is effectively designed to measure a deep conceptual understanding of Le Châtelier's Principle, not just rote memorization, by presenting a case in which the number of gas moles is equal and including distractor data. In the local context of Maros Regency, students in urban areas are likely to have greater access to supplementary learning resources, such as private tutoring or varied question banks. This exposure trains them to develop analytical and critical thinking skills. Conversely, students in rural areas with limited access may rely more on rote memorization, leaving them unprepared for non-standard questions like this.

From this finding, the practical implication to be drawn is not to simplify or eliminate Item 26, as this item is, in fact, a valid measure of Higher-Order Thinking Skills (HOTS). Instead, this result provides valuable input for education policymakers and the MGMP in Maros Regency. This finding underscores the need to equalize teaching quality by providing teachers with training across all areas to implement learning methods oriented toward conceptual reasoning and problem-solving. Thus, the instrument's quality is maintained as a reliable measurement tool, while the focus of improvement is on providing equal learning opportunities for all students, regardless of their geographical location.

Items 19, 23, & 24: DIF Based on Economic Status

Most findings came from the economic status comparison, where students from affluent families showed an advantage on complex questions requiring multi-layered reasoning and procedural mastery.

Item 19 (Voltaic Cell Notation)

A substantive analysis of Item 19 indicates a statistically significant performance difference between students from affluent and less affluent families, with the question being more difficult for students from less affluent backgrounds. However, it is important to place this finding in the proper perspective. Although statistically significant (Z Raju = 2.087; $p < 0.05$), its effect size is classified as Category 'A' (negligible). This

means that although a real performance difference was detected, its practical impact on the overall fairness of the Final School Assessment (PAS) for Chemistry instrument is considered very small and not substantial.

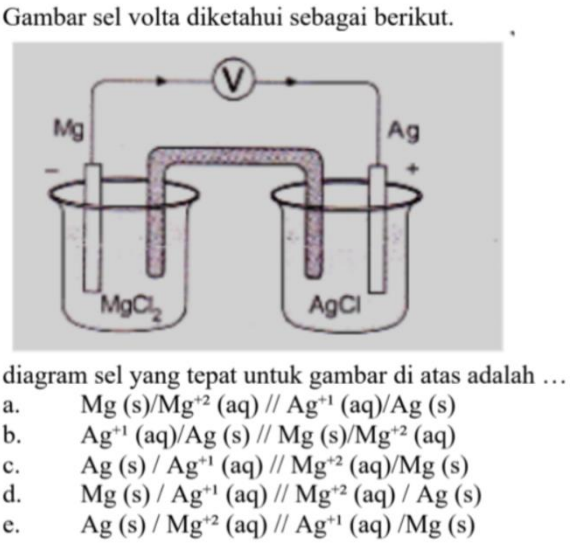


Figure 3. Indicated to Have DIF Based on Economic Status

This question requires students to translate a visual diagram of a voltaic cell into standard cell notation ($Mg(s) \mid Mg^{2+}(aq) \parallel Ag^{+}(aq) \mid Ag(s)$). Success depends on visual literacy, conceptual understanding (Anode, Cathode, Oxidation, Reduction), and mastery of the cell notation syntax. Bourdieu's Theory of Cultural Capital can explain the advantage of affluent students; they have greater access to resources (tutoring, question banks) that intensively train these three skill layers through repetition and corrective feedback.

The interpretation of this finding does not point to content bias, but rather to the high cognitive demands of the question. Item 19 specifically tests representational competence, which is the ability to translate information from one mode (a visual diagram of a voltaic cell) to another (abstract, rule-bound symbolic cell notation). Successfully answering this question heavily relies on the automaticity of procedural knowledge, which is built through intensive practice to reduce cognitive load.

In the local context of Maros Regency, this performance difference likely reflects a gap in access to supplementary educational resources. Students from higher-income families often have the opportunity to attend private tutoring, where procedures for solving complex, algorithmic problems like this are repeatedly practiced until they become automatic. In contrast, students from less affluent families who rely more on classroom learning may not get enough practice volume to achieve the same level of fluency.

The practical implication of this finding is not to revise or eliminate Item 19, as it is valid in measuring an essential skill in chemistry. Instead, this result provides important feedback for teachers and the MGMP in Maros Regency to increase in-class practice on multi-representational tasks (Istiyono et al., 2022; Nor et al., 2017). The goal is to provide all students, regardless of economic background, with more equal opportunities to build the procedural automaticity and representational competence needed to answer high-level questions.

Item 23 (Salt Hydrolysis)

The analysis of Item 23 indicates a statistically significant performance difference between students from affluent and less affluent families, with the question being more difficult for students from less affluent backgrounds. However, this finding must be understood proportionally. Although statistically significant (Z Raju = 2.567; $p < 0.05$), its effect size is classified as Category 'A' (negligible). This indicates that although a performance difference was detected, its practical impact on the overall fairness of the Final School Assessment (PAS) for Chemistry instrument is considered very small and does not substantially threaten the validity of the scores.

Garam dapat bersifat asam, basa atau netral bergantung pada asam-basa pembentuknya. Garam berikut yang bersifat asam dalam air adalah …

a. NH_4Cl
b. Na_2CO_3
c. KCl
d. CH_3COONH_4
e. NaCl

Figure 4. That is Indicated to Have DIF Based on Economic Status

This question asks to identify a salt that is acidic in nature (NH_4Cl). This requires basic memorization of strong/weak acids and bases and a conceptual understanding of hydrolysis. Affluent students are again advantaged due to access to tutoring that provides effective memorization methods and repeated practice to master the concept and avoid distractors.

The interpretation of this finding does not point to content bias but rather highlights the cognitive complexity of the question. Item 23 requires students to integrate three levels of understanding: (1) factual knowledge (memorization of strong/weak acids and bases), (2) conceptual understanding (the principle of hydrolysis), and (3) analytical skills (identifying and ignoring distractors). The difficulty faced by students from less affluent economic groups likely stems from a lack of automaticity in basic factual knowledge, which depletes their cognitive resources before they can engage in more complex conceptual reasoning.

In the local context of Maros Regency, this performance difference is very likely due to an access gap to supplementary educational resources. Students from higher-income families tend to have greater opportunities to attend private tutoring, where fundamental knowledge, like the list of strong acids and bases, is intensively drilled until it becomes automatic. Repeated exposure to varied questions also trains them to recognize patterns and common distractors. Students without similar access are more dependent on classroom understanding, which may not be sufficient to achieve the same level of fluency.

The practical implication of this finding is not to simplify Item 23, as it effectively assesses a deep, integrated understanding of chemistry. Instead, this result provides valuable feedback for teachers and the MGMP in Maros Regency. There is a need to ensure that fundamental knowledge is not only taught but also practiced to the point of automaticity for all students in the classroom (Safitri, Rosnawati, et al., 2024). Thus, the focus of improvement is on pedagogical strategies to strengthen students' knowledge foundations equitably, thereby reducing reliance on out-of-school learning resources and creating a fairer opportunity for all students to demonstrate their analytical abilities.

Item 24 (Rate Law)

The analysis of Item 24 indicates a statistically significant performance difference between students from affluent and less affluent families, with the question being more difficult for students from less affluent backgrounds. However, although statistically significant (Z Raju = 2.166; $p < 0.05$), its effect size is classified as Category 'A' (negligible). This indicates that although a performance difference was detected, its practical impact on the overall fairness of the Final School Assessment (PAS) for Chemistry instrument is considered very small and not substantial.

Pada reaksi A + B → AB diperoleh data sebagai berikut.

Percobaan	Konsentrasi Awal (M)		Laju Reaksi
	A	B	
1.	1×10^{-2}	2×10^{-2}	1×10^{-5}
2.	1×10^{-2}	4×10^{-2}	2×10^{-5}
3.	2×10^{-2}	2×10^{-2}	4×10^{-5}

Jumlah total orde reaksi dari reaksi tersebut adalah …

a. 3
b. 4
c. 5
d. 1
e. 2

Figure 5. Indicated to Have DIF Based on Economic Status

This question requires students to determine the total reaction order from a table of experimental data. This is a highly procedural question that demands the systematic and mechanical application of the comparison method. Affluent students are more likely to have mastered this procedure to the point of automaticity through intensive practice in supplementary question books or on online learning platforms.

The interpretation of this finding does not point to content bias but rather to the demands of the question, which require procedural knowledge and mathematical fluency. Item 24 requires students to systematically apply the initial rate comparison method, a multi-step algorithmic procedure involving mathematical operations with scientific notation and exponents. The difficulty students from lower-income groups face likely stems from a lack of automaticity in this procedure, which increases cognitive load and lowers self-efficacy (confidence) in mathematics.

In the local context of Maros Regency, this performance difference is very likely due to an access gap to supplementary educational resources that foster "mastery experiences." Students from higher-income families tend to have greater opportunities to attend private tutoring, where solving algorithmic problems like this is intensively practiced. This repeated exposure builds skills, speed, and confidence. Students without similar access may not achieve the same level of fluency through classroom learning alone.

The practical implication of this finding is not to eliminate or simplify Item 24, as it validly measures the ability to analyze kinetic data, which is an important competency. Instead, this result provides valuable feedback for teachers and the MGMP in Maros Regency. There is a need to strengthen procedural-based instruction and algorithmic problem-solving in the classroom. By providing adequate practice for all students, schools can help build procedural fluency and self-efficacy more equitably, thereby reducing reliance on out-of-school learning resources and creating a fairer opportunity for all students to demonstrate their abilities.

CONCLUSION

Based on the data analysis, the Chemistry PAS instrument appears generally fair. Five items showed statistically significant DIF: one based on gender, one on residential location, and three on economic status. No DIF was detected based on the school of origin. However, the effect size analysis showed that all detected DIF items fall into Category 'A' (negligible), indicating that the practical impact of the existing bias is minimal and does not significantly threaten the instrument's overall fairness or validity.

Although this chemistry assessment instrument has generally been shown to function fairly, with all indications of *Differential Item Functioning* (DIF) at negligible levels, several recommendations are suggested for continuous improvement and future evaluation practices. First, it is highly recommended that the items statistically identified with DIF (Items 11, 19, 23, 24, and 26) still undergo a qualitative review by content experts to identify potential sources of irrelevant bias, such as confusing language or context, even though their effect sizes are small. To strengthen the conclusions, a reanalysis could be considered using an alternative IRT model, such as the 2PL model, which might uncover DIF patterns masked by the 1PL model's assumptions, and applying other detection methods beyond the Raju method to increase confidence in the findings. Lastly, and most importantly, DIF analysis should be integrated as a routine procedure in every future assessment administration, especially if there are changes to item content or the demographic composition of the student population, to guarantee longitudinal assessment fairness and enhance the accountability of the evaluation system.

REFERENCES

Anizar, & Sardin. (2023). *Evaluasi pada kurikulum Merdeka dan pemanfaatan hasil penilaiannya*. Edupedia Publisher.

Amelia, R., et al. (2022). Deteksi bias gender pada instrumen evaluasi belajar kimia dengan metode Mantel-Haenszel. *Jurnal Tarbiyah*, *29*(2).

Chalmers, P. (2022). *Multidimensional item response theory*. CRAN.

Guo, H., Lu, R., Johnson, M. S., & McCaffrey, D. F. (2022). Alternative Methods for Item Parameter Estimation: From CTT to IRT (Research Report No. RR-22-12). In *ETS Research Report Series* (Vol. 12,

Issue 1). https://doi.org/10.1002/ets2.12355
Gyamfi, A., & Acquaye, R. (2023). Parameters and Models of Item Response Theory (IRT): A Review of Literature. *Acta Educationis Generalis*, *13*(3), 68–78. https://doi.org/10.2478/atd-2023-0022
Istiqlal, M., Istiyono, E., Widihastuti, Sari, D. K., Danni, R., & Safitri, I. (2024). Construction of Mathematics Cognitive Test Instrument of Computational Thinking Model for Madrasah Aliyah Students. *Nazhruna: Jurnal Pendidikan Islam*, *7*(2), 475–492. https://doi.org/10.31538/nzh.v7i2.4425
Istiyono, E., Dwandaru, W. S. B., Fenditasari, K., Ayub, M. R. S., & Saepuzaman, D. (2022). The Development of a Four-Tier Diagnostic Test Based on Modern Test Theory in Physics Education. *European Journal of Educational Research*, *12*(1), 371–385. https://doi.org/10.12973/eu-jer.12.1.371
Hardiyanti, N. (2018). *Komparasi metode dalam mendeteksi sensitivitas bias item soal USBN Kimia SMA Negeri di Kabupaten Bone dengan pendekatan teori respon item* [Tesis, Universitas Negeri Makassar].
Lerner, J. Y., McCubbins, M. D., & Renberg, K. M. (2021). The efficacy of measuring judicial ideal points: The mis-analogy of IRTs. *International Review of Law and Economics*, *68*, 1–11. https://doi.org/10.1016/j.irle.2021.106020
Magis, D., & Raiche, G. (2022). *Collection of methods to detect dichotomous differential item functioning (DIF)*. CRAN.
Nor, R., Pendidikan, E., Pascasarjana, P., & Yogyakarta, U. N. (2017). *IMPLEMENTASI ITEM RESPONSE THEORY SEBAGAI BASIS ANALISIS KUALITAS BUTIR SOAL DAN KEMAMPUAN KIMIA SISWA KOTA YOGYAKARTA Implementation of Item Response Theory for Analysis of Test Items Quality and Students ' Ability in Chemistry*. *2*(1), 1–12.
Pratama, D., dkk. (2021). Analisis differential item functioning (DIF) pada skala sikap moderasi beragama siswa. *Proyeksi*, *18*(1), 116–131.
Pristiwaluyo, T., & Syamsuddin, S. (2021). Development of an instrument for teachers' attitudes towards academic supervision performed by supervisors in schools of special education. *Journal of Educational Science and Technology (EST)*, *7*(1), 40–49.
Priyadi, S. (2024). Differential item functioning (DIF): Sebuah analisis bibliometrik. *Edukatif: Jurnal Ilmu Pendidikan*, *6*(4), 3975–3989.
Rust, J., Kosinski, M., & Stillwell, D. (2020). Modern Psychometrics. In *Modern Psychometrics*. https://doi.org/10.4324/9781315637686
Safitri, I., & Ansyari, R. (2024). The Calibration of Science Achievement Test Based on Integrated Islamic Curriculum. *8th International Conference on Education and Multimedia Technology (ICEMT)*, 329–336. https://doi.org/10.1145/3678726.3678727
Safitri, I., Lestarani, D., Imtikhanah, R. D. N. W., Akbarini, N. R., Sari, M. W., Fitrah, M., & Hapsan, A. (2024). *TEORI PENGUKURAN DAN EVALUASI*. CV. Ruang Tentor.
Safitri, I., Rosnawati, R., & Ansyari, R. (2024). Estimasi Kesalahan Pengukuran dalam Penilaian Sidang Skripsi: Generalizability Theory Analysis. *Afeksi: Jurnal Penelitian Dan Evaluasi Pendidikan*, *5*(1), 162–168. https://doi.org/https://doi.org/10.35672/afeksi.v5i1.220
Samritin, S. (2022). Identifikasi muatan differential item functioning pada data Ujian Nasional Matematika. *Journal on Education*, *4*(4), 1675–1684.
Setiawan, A. (2020). Pendeteksian DIF pada perangkat tes objektif penilaian akhir semester IPA dengan menggunakan permodelan Rasch. *PSEJ (Pancasakti Science Education Journal)*, *5*(2), 23–29.
Thiyagarajan, A., James, T. G., & Marzo, R. R. (2022). Psychometric properties of the 21-item Depression, Anxiety, and Stress Scale (DASS-21) among Malaysians during COVID-19: a methodological study. *Humanities and Social Sciences Communications*, *9*(1), 1–8. https://doi.org/10.1057/s41599-022-01229-x
Yin, R. K. (2018). *Case study research: Design and methods* (6th ed.). SAGE Publications.
Wahyuni, A. (2022). Detection of gender bias using DIF (differential item functioning) analysis on the item test of the school examination in Yogyakarta. *Jurnal Evaluasi Pendidikan*, *13*(1).