Test Item Bias Analysis Using Differential Item Functioning (DIF): A Mantel-Haenszel Chi-Square Statistics Approach

Arlene Nisperos Mendoza, Pangasinan State University, Philippines

The Paris Conference on Education 2025
Official Conference Proceedings

Abstract

This study highlights the importance of implementing Differential Item Functioning (DIF) analysis to assess the fairness and validity of educational measures. The analysis examines possible test item biases against certain groups of test-takers based on factors like age, sex, socio-economic status, and school type. Utilizing the Mantel-Haenszel Chi-Square Statistic, the study identified biased test items, with over one-third exhibiting bias, consequently compromising the assessment's fairness and validity. The findings demonstrated that age, sex, socioeconomic status, and the type of educational institution exerted a discernible influence on the disparities observed in students' performance on the examination. Moreover, it was ascertained that age played a particularly significant role in these variations. Removing potentially biased items resulted in a more equitable and valid assessment, emphasizing the importance of identifying potential biases to enhance the test's quality and reliability, ultimately contributing to the improvement of educational assessment.

Keywords: differential item functioning (DIF), Mantel-Haenszel chi-square statistics, mathematics achievement test, item bias, item reliability, test validity

iafor
The International Academic Forum
www.iafor.org

Introduction

In the realm of educational assessment, it is of utmost importance to ensure that tests are fair and equitable for all examinees. According to Wetzel and Böhnke (2017), the responses observed from individuals should solely depend on their inherent abilities and not be influenced by external factors like gender. To address the historical disparities in test-taking populations caused by systemic inequality, statistical and psychometric tools can be employed to identify and eliminate test items that perpetuate the problem, as pointed out by Lucey and Saguil (2020). One such powerful tool available for this purpose is Differential Item Functioning (DIF) analysis (Wetzel & Böhnke, 2017).

DIF analysis has long been recognized as a fundamental aspect of educational assessment, particularly in the domain of large-scale assessments like the Trends in Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA) (Chen & Jin, 2018; Stark et al., 2006). The Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) emphasize the importance of validity as the primary consideration in test development and usage. To ensure a meaningful and appropriate interpretation of test scores, incorporating evidence from various sources, including DIF analyses, is recommended (AERA, APA, & NCME, 2014; Wu et al., 2018).

DIF occurs when a test item displays a difference in the probability of correctly answering the item among individuals from different groups, even when matched on the underlying latent trait (Wetzel & Böhnke, 2017). Typically, this difference is based on demographic attributes such as gender, ethnicity, or language. The presence of a significant number of items with DIF poses a serious threat to the construct validity of tests and the inferences drawn from test scores obtained from items with and without DIF. Karami (2012) notes that if the factor contributing to DIF is not related to the construct being tested, the test results become biased. In other words, if DIF is not taken into account, it can distort the test scores and lead to invalid conclusions about examinees' performance on the test.

In recent times, there has been growing concern and criticism surrounding standardized testing, particularly regarding test inequality. This criticism suggests that educational assessments used both in North America and globally exhibit systematic discrimination against marginalized groups of students, such as Black, Latino, and low-income students (Koljatic et al., 2021; Sireci, 2021), as indicated by Miranda (2020). As a response to this problem, the University of California System has decided to discontinue the use of the ACT and SAT for admission purposes (Moskowitz, 2022; Rio, 2021).

An incident in Nueva Vizcaya involving a misrepresented document about Igorots has raised questions about the Philippine government's duty to provide quality, equitable, culture-based, and comprehensive basic education (Department of Education [DepEd], 2021). The DepEd (2021) strictly enforces a zero-tolerance policy on discrimination of any kind. Once the error was found, field offices swiftly withdrew the document, preventing learners from accessing it.

Jones (2019) warns that the presence of differential item functioning (DIF) can introduce bias in assessing group differences and compromise research outcomes and risk factors. Similarly, Garcia et al. (2021) investigated the psychometric properties of the Beck Anxiety Inventory

(BAI) across various demographic variables in a multi-ethnic cohort. While the BAI has proven effective for measuring anxiety symptoms in Hispanic/Latino Americans and Non-Hispanic/Latino Americans (Bardhoshi et al., 2016), further validation of its cross-cultural applicability is recommended for improved measurement accuracy.

Similarly, Almarabheh and Alshammari (2020) identified sex-related differential item functioning (DIF) in Raven's Standard Progressive Matrices (SPM) Test. Their study revealed biased items against female performance, highlighting the need for additional analysis using item response theory-based techniques like logistic regression, simultaneous item bias test (SIB), or IRT-likelihood ratio (IRT-LR) methods to confirm the findings.

Despite the importance of DIF analysis in evaluating biases in testing, its complexity has limited its adoption among researchers who are less mathematically inclined (Karami, 2012). The intricacies of DIF analysis stem from the underlying statistical and psychometric concepts involved, typically requiring advanced statistical techniques and a solid understanding of the theoretical framework. This can pose challenges for researchers with limited mathematical or psychometric expertise.

The limited adoption of DIF analysis among less mathematically oriented researchers has significant implications for the development and implementation of testing practices. Ensuring fairness and equity in testing practices for individuals, irrespective of their demographic characteristics, is crucial. Therefore, efforts should be made to enhance the accessibility of DIF analysis and develop simpler and more user-friendly procedures that researchers with varying levels of mathematical and psychometric expertise can easily understand and implement.

The objective of the current study was to develop a reliable, valid, and fair test by detecting bias in test items. To achieve this goal, the researcher employed the Mantel-Haenszel (MH) Chi-Square Statistics to identify biased items in the Mathematics Achievement Test. This method has been found to be effective in detecting bias in dichotomously scored tests.

In a study by Rustam et al. (2019), the Mantel-Haenszel (MH) method outperformed the standardization method in detecting DIF for samples of 400 and 2000. The authors acknowledged that the standardization method may be suitable for smaller sample sizes or imbalanced focus groups. However, the superiority of either method should not be assumed.

Similarly, Al-Batosh and Qur'an (2018) used the MH method to investigate DIF in assessment tools for higher education quality in Jordan, focusing on different academic colleges. Their findings revealed bias in favor of Science faculty students, disadvantaging Education and Arts faculty students. Additionally, DIF significantly impacted the internal construction validity indicators of the assessment tool.

Moreover, the current study assessed the effects of eliminating biased items on test quality measures, including content and concurrent validity, as well as internal consistency reliability. The findings hold considerable significance for the creation and implementation of equitable and dependable assessments, particularly in contexts where testing carries high stakes.

The identification and elimination of biased items play a pivotal role in the creation of fair and equitable assessments, particularly in high-stakes testing environments. In this regard, the present study holds immense potential to make a profound contribution to the field of educational research, specifically in the realm of test development. Test experts, developers, and educators stand to gain valuable insights from this study. Firstly, they can acquire a deep understanding of the applicability of Differential Item Functioning (DIF) detection methods. Secondly, they can recognize the validity of DIF methods in identifying biased test items based on students' diverse characteristics, such as age, sex, socioeconomic status, and school type. Thirdly, they can utilize DIF methods to construct assessments that are both valid and equitable. Lastly, they can employ DIF methods to refine their assessment instruments, thus augmenting the precision and impartiality of their tests. In summary, this study offers a valuable framework for enhancing the quality and equity of educational assessments, ultimately benefiting both students and teachers.

Literature Review

This research is grounded on Measurement Invariance (MI), a fundamental psychometric concept that guarantees the interpretation of scores similarly across groups. DIF points to MI violations at the item level, which implies bias. The research employs the Mantel-Haenszel (MH) approach, one of the most popular methods of detecting DIF.

Measurement Invariance

Measurement invariance, or construct equivalence, ensures that a test measures the same underlying trait across different groups. DIF indicates that people with the same trait level from different groups have different probabilities of answering an item correctly.

The study's aim to "assess the fairness and validity of educational measures" and identify "possible test item biases" directly addresses the principle of measurement invariance. The presence of DIF, as detected by the Mantel-Haenszel method, serves as empirical evidence that the mathematics achievement test is not invariant across the specified demographic groups.

The Mantel-Haenszel (MH) Method Within Classical Test Theory (CTT)

The MH method, based on Classical Test Theory (CTT), is a non-parametric technique where an observed score equals a true score plus error. In DIF analysis, CTT helps create "matching" variables, like total scores, as proxies for ability, allowing comparison of item performance between groups while controlling for overall proficiency.

The study's explicit choice of the MH Chi-Square Statistic aligns with its advantages for dichotomously scored tests and its robustness, as highlighted by Rustam et al. (2019). While other methods exist, this study leverages the MH approach's practical utility for identifying bias in a straightforward and interpretable manner.

Test Fairness and Educational Equity

Beyond statistical invariance, this framework emphasizes the ethical obligation for test fairness. Fair assessment demands tests measure intended traits without unfairly disadvantaging subgroups based on irrelevant traits. The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) states that validity—support for score interpretation—is tied to fairness.

The research supports social and educational equity. It references concerns about "systematic discrimination against marginalized groups" in standardized tests, leading to policy changes like the University of California System dropping ACT/SAT. The study illustrates a psychometric approach to addressing such inequities, especially in the diverse Philippine education system with reports of cultural insensitivity (DepEd, 2021). Findings on factors like age, sex, socioeconomic status, and school type reveal areas where psychometric vigilance can foster fairer educational outcomes.

Enhancing Test Quality and Accessibility

The framework shows that addressing DIF improves test validity and reliability. Removing biased items strengthens construct validity by ensuring the test measures the intended construct consistently across groups. It also boosts internal consistency by ensuring items contribute cohesively to the overall score without extraneous variance from group-specific factors.

Acknowledging Karami's (2012) note on the complexity of DIF analysis limiting its use among less mathematically inclined researchers, this study contributes to the discussion on making psychometric tools more accessible. By demonstrating the Mantel-Haenszel Chi-Square Statistic, it provides a practical example for test experts, developers, and educators, encouraging the adoption of DIF analysis for improving assessments. This shows that even without advanced IRT knowledge, robust DIF detection is possible.

Methodology

Research Design

This research undertook a comprehensive exploration using a descriptive research design, employing an achievement test in Mathematics to assess student performance. To analyze the test items, the researcher employed the Differential Item Functioning (DIF) method, specifically the Mantel-Haenszel Chi-Square Statistic approach. The DIF analysis focused on investigating potential variations among different groups, encompassing factors such as age, sex, socioeconomic status, and school type. Through this rigorous examination, the study sought to uncover valuable insights into the presence of any differentials in item performance, shedding light on the potential influence of various demographic variables on test outcomes.

Respondents of the Study

The examination was administered to all students in the higher education program for Secondary Education, with a focus on Mathematics. These students were from State Universities and Colleges, as well as some selected Higher Education Institutions in Region I. They were specifically chosen because they successfully completed the Calculus course, which was an essential part of their required curriculum within their chosen academic field.

Data Gathering Instrument

In this research study, a structured questionnaire was developed for the purpose of collecting information on several key demographic factors among the student population, including age, sex, socioeconomic status, and school type. This data was subsequently used to identify and detect any potential biases present in the questionnaire items.

Furthermore, to gather quantitative data regarding the academic performance of the students, an academically rigorous achievement test focusing on the subject of Calculus was administered. This meticulously designed test consisted of a robust set of 100 multiple-choice items, crafted to encompass a wide range of crucial concepts. These concepts encompassed Functions (8 items), Limits and Continuity (27 items), Derivatives (31 items), and Analysis of Functions and their Graphs (34 items), ensuring comprehensive coverage of the subject matter.

Data Gathering Procedure

With the objective of developing a valid and equitable mathematics test, the constructed test underwent a rigorous evaluation by a panel of mathematics experts, followed by extensive field testing among mathematics-specializing students. After successful content validation, the test was administered to a representative sample of students enrolled in the program of interest across multiple state universities, colleges, and selected higher education institutions within the region. The resulting test scores were analyzed, considering key demographic variables such as age (17 and below or 18 and above), sex (male or female), socioeconomic status in terms of gross monthly income (PHP 8,000.00 and below or above PHP 8,000.00), and school type (public or private).

These diverse groups formed the basis for item analysis, employing the DIF method. This method enabled the identification of potential performance disparities among subgroups, offering valuable insights into factors influencing academic achievement. The detection of DIF guided the elimination and improvement of test items. The revised version of the test subsequently underwent further tests of validity and reliability using established statistical methods.

Statistical Treatment of Data

Detection of Bias Items Using the Mantel-Haenszel Chi-Square Statistic Approach

This study aimed to identify bias in test items using MH Statistics to detect DIF. By analyzing odds ratios across subgroups, significant performance differences indicated item bias. A follow-up investigation sought the bias sources.

The computation of the MH Statistic commenced with the determination of the probabilities of correct and incorrect responses for both the focal and reference groups. This was followed by assessing the relative likelihood of each group answering an item correctly. The overall measure of DIF was obtained by aggregating the odds ratios across all ability levels and normalizing them based on the number of ability levels. The resulting index is known as the Mantel-Haenszel odds ratio, denoted by α_{MH}, which is commonly transformed using the formula: $\beta_{MH} = \ln \alpha_{MH}$ (Karami, 2012).

According to Wiberg (2007), a negative value of β_{MH} indicated the presence of DIF favoring the focal group, while a positive value indicated DIF favoring the reference group. In some cases, β_{MH} was further recalibrated into $MHD = -2.35 \ln \alpha_{MH}$. The Mantel-Haenszel Delta (MHD) serves as an indicator of the degree of DIF. As noted by Karami (2012), a positive

MHD value indicated that the test item presented greater challenges for the reference group, whereas a negative value indicated that the focal group experienced greater difficulty with the item.

The Mantel-Haenszel Differential Item Functioning (MH DIF) analysis utilized a Chi-Square Statistic to assess item bias. This statistic was compared to a critical value of 3.8415 at a significance level of 0.05, with one degree of freedom, serving as a detection threshold for potentially biased items. Items exceeding the threshold with an MH Chi-Srequa Statistic value were flagged as displaying Differential Item Functioning (DIF) and underwent further analysis to identify the source of bias.

Pedrajita (2015) proposed a classification system to categorize the degrees of DIF in test items into three levels: A, B, and C. This system aims to avoid identifying items with statistically significant DIF that are practically trivial. The categories are defined as follows: Category A: Items flagged as A show negligible amounts of DIF, with the absolute value of the Mantel-Haenszel Delta (MHD) significantly differing from 0 but smaller than 1; Category B: Items identified as B exhibit moderate levels of DIF, with MHD values significantly differing from zero and either not significantly greater than 1.0 or smaller than 1.5 in absolute value; Category C: Items falling into the C category display large amounts of DIF, with the absolute value of the MHD being greater than 1.5 or significantly different from 1.0. The table below provides a summary of these categories:

Table 1
Detection Threshold and Effect Size of Mantel-Haenszel Chi-Square Statistics DIF Detection Method

Detection Threshold	Effect Size	Code	Scale Used
3.8415	0.0 – 1.0	A	Delta Scale
	1.0 – 1.5	B	
	> 1.5	C	

Validity and Reliability of the Achievement Test

The study employed various methods to evaluate the validity and reliability of the test. Firstly, a factor analysis approach was utilized to assess the construct validity of the test, examining the interrelatedness of its factors to demonstrate its unidimensionality.

Concurrent validity was established by analyzing the relationship between predictors, such as examinees' scores in the Calculus achievement test, and the criterion variable, which was their grade point average (GPA). This relationship was quantified using the Pearson Product Moment correlation coefficient, commonly known as a validity coefficient (Pedrajita, 2015).

To evaluate the content validity of the test, a content validity index (CVI) was calculated using a 5-point rating agreement scale. Content experts provided ratings of scale relevance, ensuring that the test adequately represented the intended content domain.

The internal consistency reliability of the revised test was assessed using the KR-20 formula, designed for dichotomous test items. It measured item heterogeneity and test consistency.

This evaluation helped the researcher verify the effectiveness of revisions and the test's overall reliability.

Results and Discussion

This section presents research findings using the MH Chi-Square Statistic to identify biased items across student variations in age, sex, socio-economic status, and school type. It also highlights significant findings regarding the validity and reliability of the revised achievement test.

Detection of Bias Items Using Mantel-Haenszel Chi-Square Statistic Approach

The results of the MH analysis were presented in Tables 2, 3, 4, 5, and 6, accompanied by Figures 1, 2, 3, and 4, providing a visual representation of the findings.

Based on Age Differences

The study analyzed measurement invariance by age, with results in Table 2 and Figure 1. Using DIF detection, 23 items showed bias: four (8, 16, 37, 96) toward the group 17 and below, and 19 (18 and above) toward the reference group.

Table 2
Biased Items With Significant DIF Across Age Using MH

Item No.	MH χ^2 Statistic	p-value	α_{MH}	MHD	Potentially Biased Groups
8	5.2129*	0.0224	Inf	-Inf ^{C}	Focal
10	11.9951**	0.0005	0.0905	5.6459 ^{C}	Reference
11	4.2082*	0.0402	0.1588	4.3238 ^{C}	Reference
13	4.0750*	0.0435	0.2468	3.2882 ^{C}	Reference
15	10.1538**	0.0014	0.0848	5.7983 ^{C}	Reference
16	12.8489**	0.0003	Inf	-Inf ^{C}	Focal
19	5.6008*	0.0180	0.1640	4.2492 ^{C}	Reference
22	15.8631**	0.0001	0.0503	7.0264 ^{C}	Reference
34	8.2728**	0.0040	0.0590	6.6501 ^{C}	Reference
37	6.1449*	0.0132	Inf	-Inf ^{C}	Focal
44	6.3181*	0.0120	0.1825	3.9977 ^{C}	Reference
52	6.8628**	0.0088	0.1736	4.1145 ^{C}	Reference
53	11.9554**	0.0005	0.0928	5.5874 ^{C}	Reference
57	9.9643**	0.0016	0.0447	7.3011 ^{C}	Reference
58	6.2843*	0.0122	0.1174	5.0350 ^{C}	Reference
61	7.4552**	0.0063	0.0729	6.1546 ^{C}	Reference
66	4.2206 *	0.0399	0.0662	6.3816 ^{C}	Reference
69	11.1771**	0.0008	0.0550	6.8156 ^{C}	Reference
72	14.0188**	0.0002	0.0486	7.1054 ^{C}	Reference
74	16.8953**	0.0000	0.0626	6.5101 ^{C}	Reference
77	5.1305*	0.0235	0.1332	4.7365 ^{C}	Reference
81	6.2241*	0.0126	0.0772	6.0193 ^{C}	Reference
96	4.7843*	0.0287	10.4840	-5.5221 ^{C}	Focal

Legend: * significant at α = .05 ** significant at α = .01; C – large DIF

Among the items analyzed, it was found that Item 74 showed the highest level of DIF, putting the reference group at a disadvantage. This item challenged students to understand the first derivative of an exponential function, which may have been particularly difficult for individuals aged 18 and older. Conversely, Item 13 had the lowest DIF estimate, as indicated by the MH Chi-Square Statistic. This item involved inverse functions and asked students to find the value of a composite function using its inverse, making it somewhat less challenging for the test takers.

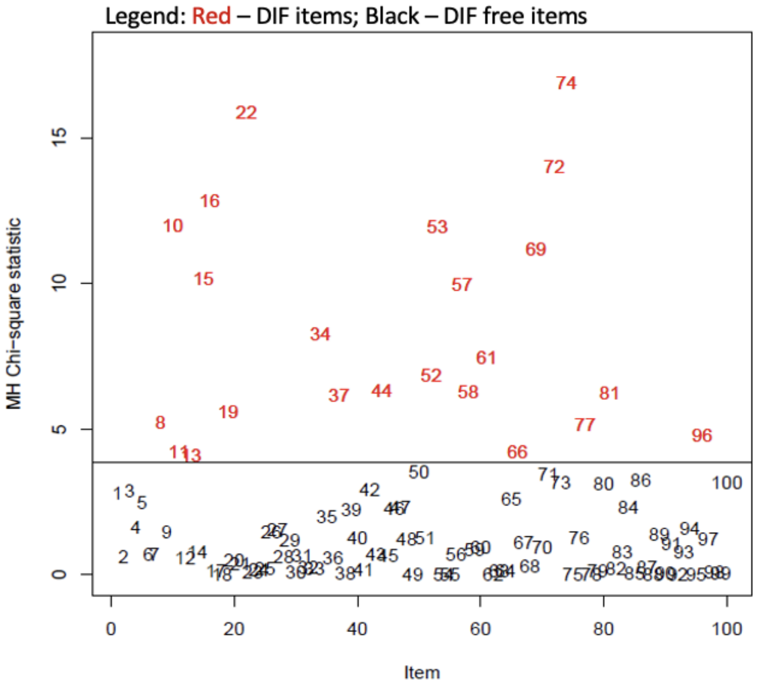
The analysis yields a compelling outcome, revealing a notable discrepancy in the number of DIF items that may exhibit bias against the reference group, as elegantly demonstrated in the accompanying table. These intriguing findings suggest the possibility that individuals aged 17 and below possess a heightened capacity for memory retention, granting them a distinct advantage in recalling the intricacies covered in Calculus. In contrast, those aged 18 and above may benefit from engaging in additional review of the subject matter addressed in the identified DIF items, as these particular questions could hold significant importance for their impending board examination.

This phenomenon is believed to be associated with the ongoing growth and maturation of the brain, coupled with the intricate formation of neuronal connections during this crucial developmental stage of life (Eichenbaum, 2017; Keresztes et al., 2017).

Furthermore, it is noteworthy that all the identified DIF items have been classified under the category "C," signifying a substantial effect and necessitating rigorous revision or potential replacement. In contrast, the remaining items not featured in the table have been assigned classifications of "A" and "B" and have not exhibited any detectable DIF. Nevertheless, these items are still visible and traceable in Figure 1, providing valuable insights into their performance characteristics.

Figure 1

Item Bias Detection Using MH Statistics Across Age



The graphical representation in Figure 1 depicts the deviation of each item's MH Chi-Square Statistics from the critical value. This visual aid serves to provide support for the obtained outcomes, as exemplified in Table 2.

The study highlights tailoring teaching and assessments to diverse student abilities across ages, improving fairness and inclusivity while fostering better learning. Adjustments for age differences ensure unbiased testing, recognizing each group's unique needs and promoting equitable evaluation.

Based on Sex Differences

The present analysis provides significant insights into sex differences in the MH analysis, as shown in Table 3. Importantly, only items 37 and 72 showed potential bias against the reference and focal groups, respectively.

Table 3
Biased Items With Significant DIF Across Sex Using MH

Item No.	MH χ^2 Statistic	p-value	α_{MH}	MD	Potentially Biased Groups
37	4.1693*	0.0412	0.3531	2.4461^C	Reference
72	5.1985*	0.0226	3.7984	-3.1363^C	Focal

Legend: * significant at $\alpha = .05$; C – large DIF

The analysis has uncovered noteworthy insights regarding item 37 in the test, which evaluates students' proficiency in identifying properties of the graph of y = arctan x beyond mere graphing skills. Intriguingly, the reference group exhibited challenges in answering this item, hinting at a potential knowledge gap in graphing the function. These findings suggest the possibility of lower vigilance among male examinees regarding the graphical representation of mathematical functions, as evidenced by the results. Further research and investigation are warranted to comprehensively comprehend and address these potential disparities.

On the other hand, item 72 pertains to the determination of $\frac{d}{dx}\left(\frac{f+g}{h}\right)$, considering that *f*, *g*, and *h* are differentiable functions of *x*. The focal group exhibited difficulty in tackling this item, potentially attributed to its emphasis on both differentiable functions and the rules of differentiation. This observation implies that female examinees may encounter challenges when it comes to generalizing mathematical rules and computations.

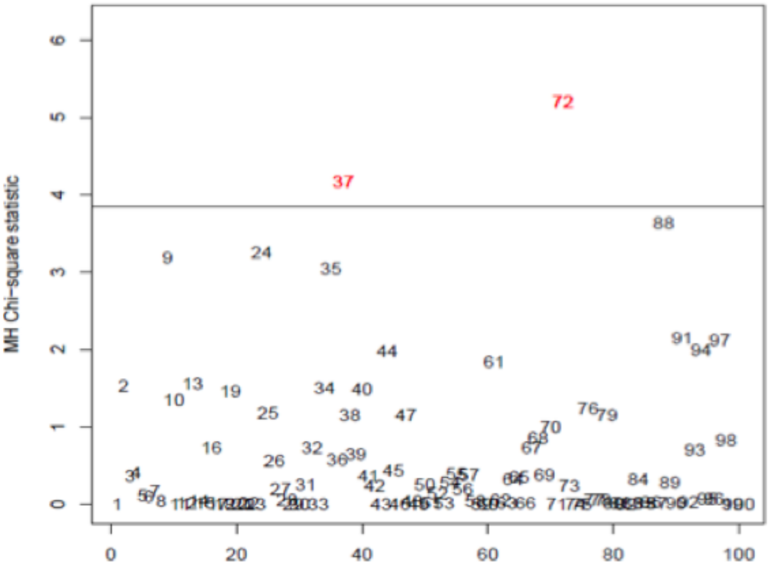
Both DIF items are categorized as "C," indicating potential for significant impact on the test. These items were also identified in MH analysis based on age differences, confirming their bias susceptibility. Findings are supported by Figure 2.

Dale et al. (2025) noted that DIF can appear "despite differences in content areas and bias directions," which supports the study’s findings of different patterns (males struggling with one item, females with another). This improves the generalizability of the results by showing sex-based DIF, indicating that DIF related to sex exists across various fields.

This study, as well as Dale et al. (2025), emphasizes the need for ongoing vigilance in test development. The consistent detection of DIF in mathematics and medical education demonstrates that test items can unintentionally disadvantage certain groups, even when overall ability is similar. Dale et al. (2025) stated that the validity of score-based inferences, especially for group comparisons, depends on test items functioning equally across different groups.

Figure 2

Item Bias Detection Using MH Statistics Across Sex



Legend: Red – DIF items; Black – DIF free items

Based on Socioeconomic Status Differences

Table 4 shows the results from the Mantel-Haenszel (MH) analysis, which examined how differences in socioeconomic status affect the test items. In this analysis, two items, specifically item 47 and item 86, were identified as showing DIF with a severity rating of "C." This rating indicates a significant level of DIF for these items.

Table 4

Biased Items With Significant DIF Across Socio-Economic Status Using MH

Item No.	MH χ^2 Statistic	p-value	α_{MH}	MHD	Potentially Biased Groups
47	4.1455*	0.0417	2.5450	-2.1952 ^{C}	Focal
86	8.0617 **	0.0045	0.2179	3.5806 ^{C}	Reference

Legend: * significant at α = .05 ** significant at α = .01; C – large DIF

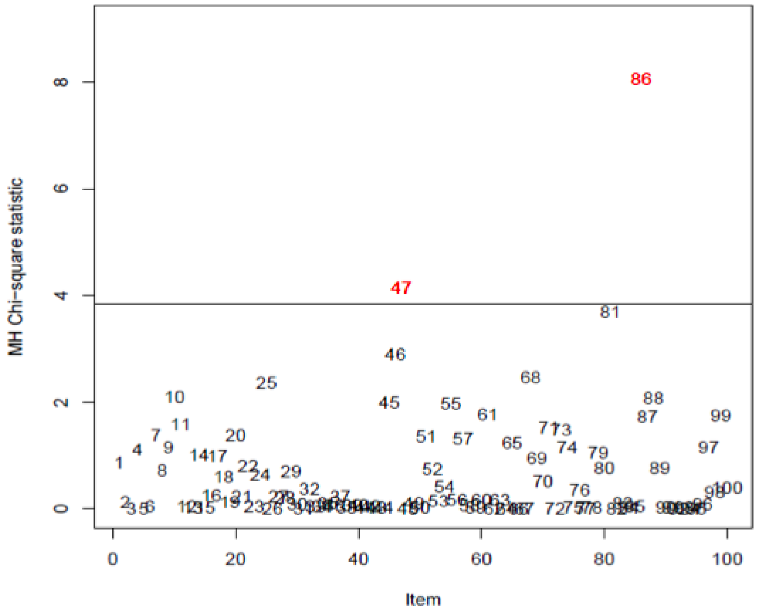
Item 86 exhibited a substantial Mantel-Haenszel DIF value, implying a potential bias against the reference group comprised of students with a monthly income exceeding Php 8,000. Students belonging to this group encountered greater challenges in achieving success on this particular item. Furthermore, the item demonstrated a significantly high level of significance, underscoring a noteworthy correlation between students' likelihood of succeeding on the item and their socio-economic status differences.

In contrast, item 47 placed the focal group, encompassing students with a monthly gross income below Php 8,000, at a disadvantage. This item assessed their comprehension of the behaviors exhibited by the graph of a given function within a specific interval of x, encompassing intricate concepts concerning limits and continuity of a function. The complexity inherent in these concepts may have induced confusion among the students.

Similar to item 86, item 47 also demonstrated a significant level of significance, signifying a meaningful relationship between students' likelihood of achieving success on the item and their socioeconomic disparities (Tan, 2024). Therefore, it is advisable to revise or potentially remove these items to mitigate bias in evaluating students' performances with respect to their socioeconomic backgrounds.

Figure 3

Item Bias Detection Using MH Statistics Across Socio-Economic Status



Legend: Red – DIF items; Black – DIF free items

Figure 3 visually reaffirms Table 4, showing black items as those without DIF in the Mantel-Haenszel analysis for socioeconomic disparities. This indicates that socioeconomic differences significantly influence students' overall test performance (Tan, 2024).

Based on School Type Differences

Table 5 presents the findings derived from the MH analysis, meticulously assessing the implications of variances in school type on the test items.

Table 5
Biased Items With Significant DIF Across School Type Using MH

Item No.	MH χ^2 Statistic	p-value	α_{MH}	MHD	Potentially Biased Groups
33	4.2754*	0.0387	4.2113	-3.3788 ^{C}	Focal
43	5.5413 *	0.0186	6.7612	-4.4913 ^{C}	Focal
44	5.5817 *	0.0181	5.4396	-3.9802 ^{C}	Focal
55	4.4308 *	0.0353	3.7161	-3.0848 ^{C}	Focal
59	4.4721 *	0.0345	14.2667	-6.2461 ^{C}	Focal

Legend: * significant at $\alpha = .05$; C – large DIF

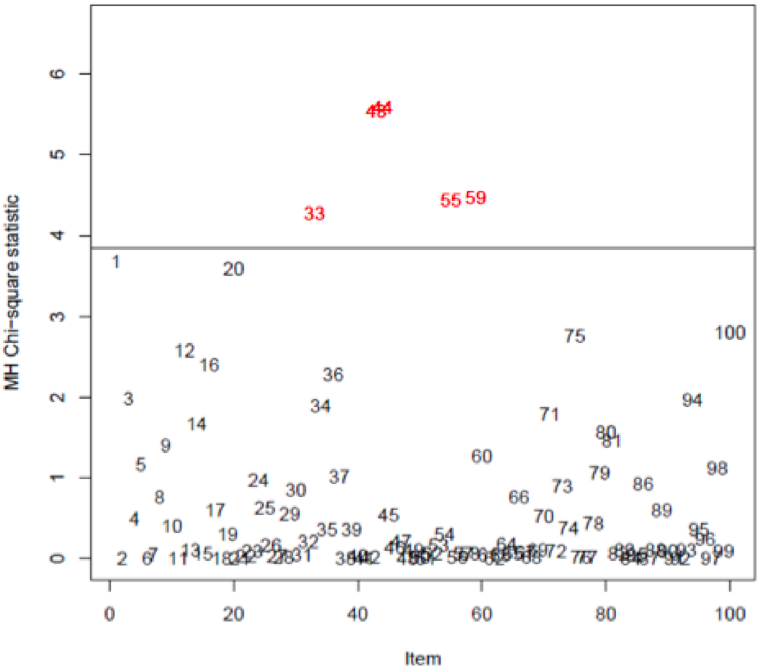
The analysis presented in Table 5 unveils compelling insights regarding the DIF of five specific test items (namely, items 44, 43, 59, 55, and 33) in relation to school type differences. These items manifest notable instances of severe DIF, signified by their classification within the "C" category, indicating a potential bias against the focal group. Particularly noteworthy is the observed difficulty experienced by students enrolled in public higher education institutions (HEIs) when attempting to excel in these items, suggesting a potential bias targeting this specific student cohort. Such biases may emanate from diverse factors, encompassing disparities in curriculum, instructional methodologies, or student backgrounds. These findings distinctly underscore the paramount significance of diligently identifying and mitigating the origins of DIF within educational assessments to ensure impartial and accurate evaluations of student performance.

Item 44 exhibits the highest DIF among the five items influenced by school type, indicating it does not function equally for students from different schools. It involves understanding a rational function's curve and properties, which requires algebraic and graphical skills. Students, especially from public HEIs with less proficiency, may struggle.

The implications of the findings suggest that incorporating computer-aided materials and projectors to present real graphs of functions can significantly enhance students' comprehension of the subject matter. By serving as valuable visual aids, these teaching tools foster better understanding and retention among learners (Pope, 2023).

Figure 4 shows items deviating from the detection threshold, suggesting they may function differently among student groups and introduce bias.

Figure 4
Item Bias Detection Using MH Statistics Across School Type



Legend: Red – DIF items; Black – DIF free items

Validity and Reliability of the Revised Achievement Test

Based on the DIF, validity, and reliability results, the achievement test was revised to 50 items covering the four Calculus subtopics. Table 6 shows the revised test's validity and reliability indices.

Table 6
Validity and Reliability Test of the Revised Achievement Test

Measures	Coefficient	Description
Construct Validity	0.667	Good
Concurrent Validity	0.159	Significant
Content Validity	0.9793 and 0.8965	Acceptable
Internal Consistency Reliability	0.822	Good

The construct validity coefficients have shown that the revised version of the Achievement Test exhibits strong psychometric properties, confirming its effectiveness as a reliable assessment tool. Additionally, the results indicate that the selected test items in the revised version align well with a single underlying dimension. Moreover, the concurrent validity coefficient provides evidence of a positive and statistically significant correlation between the test score and the grade point average in Calculus I. This supports the validity of the revised test, confirming its capacity to measure the intended construct. Furthermore, the content validity indices of the revised test meet the acceptable threshold, and expert evaluations further support its validity.

On the other hand, the data in Table 6 show a reliability coefficient above 0.8 for the revised version of the test, indicating strong internal consistency and a dependable scale for measuring students' performance.

The importance of these findings is underscored by their role in confirming the reliability of the revised test as a valid tool for assessing Calculus proficiency among students. This evaluation, in turn, becomes a valuable resource in making well-informed decisions concerning student progression, curriculum enhancement, and instructional planning. The data obtained from this assessment can serve as a powerful instrument for guiding instructional methodologies, pinpointing areas of excellence and deficiency, and implementing focused interventions aimed at enhancing student learning outcomes.

Conclusions

The application of the Mantel-Haenszel Chi-Square Statistics technique to assess Differential Item Functioning (DIF) has yielded valuable insights into test item performance and potential biases linked to age, gender, socioeconomic position, and school type.

It also highlighted the significance of construct validity, concurrent validity, content validity, and reliability analyses in assessing test item quality. The construct validity coefficients confirmed the overall success of the redesigned Achievement Test, suggesting its capacity to appropriately assess the desired construct. The concurrent validity coefficient found a favorable and substantial link between the test score and the grade point average in Calculus I, bolstering the redesigned test's validity. The content validity indices suggested that the test items were suitable, and the reliability coefficient confirmed the test's internal consistency and dependability.

These findings highlight the need of using robust statistical approaches to evaluate the quality of test items. Educators and researchers may acquire useful insights into the performance and biases of test questions by using this technique, guiding choices about curriculum creation, instructional planning, and student growth.

Acknowledgements

The researcher expresses heartfelt gratitude to Divine Providence for wisdom and blessings, making her desire possible. She also thanks the Bachelor of Secondary Education College students majoring in Mathematics and their Professors in Calculus I across Higher Education Institutions in Region 1 for their cooperation during data gathering.

Declaration of AI-Assisted Technology in the Writing Process

The researcher would like to state that she used Grammarly to improve the language of the research content.

References

Al-Batosh, A. Y., & Qur'an, M. F. A. (2018). Examining the differential item functioning in students' assessment instruments for quality of higher education in Jordan according to academic collage using Generalized Mantel-Haenszel method. *Journal of Al-Quds Open University for Educational & Psychological Research & Studies, 8*(23). https://digitalcommons.aaru.edu.jo/jaqou_edpsych/vol8/iss23/12

Almarabheh, A., & Alshammari, S. (2020). Detection of Sex-Related differential item functioning in Raven's standard progressive matrices test using the Mantel-Haenszel method. *Journal of Education and Practice*. https://doi.org/10.7176/jep/11-29-10

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), (2014). Standards for educational and psychological testing. https://eric.ed.gov/?id=ED565876

Bardhoshi, G., Duncan, K., & Erford, B. T. (2016). Psychometric meta-analysis of the English version of the beck anxiety inventory. *Journal of Counseling & Development, 94*(3), 356–373. https://doi.org/10.1002/jcad.12090

Chen, H., & Jin, K. (2018). Applying logistic regression to detect differential item functioning in multidimensional data. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.01302

Dale, E. D., Abulela, M. A. A., Jia, H., & Violato, C. (2025). Are medical school preclinical tests biased for sex and race? A differential item functioning analysis. *BMC Medical Education, 25*(1). https://doi.org/10.1186/s12909-024-06540-6

Department of Education (DepEd). (2021, February 9). DepEd reiterates zero tolerance vs. discrimination after erroneous document on Igorots surfaces. https://www.deped.gov.ph/2021/02/09/on-addressing-discrimination/

Eichenbaum, H. (2017). Memory: organization and control. *Annual Review of Psychology, 68*, 19–45. https://doi.org/10.1146/annurev-psych-010416-044131

Garcia, J. M., Gallagher, M. W., O'Bryant, S. E., & Medina, L. D. (2021). Differential item functioning of the Beck Anxiety Inventory in a rural, multi-ethnic cohort. *Journal of Affective Disorders, 293*(1), 36–42. https://doi.org/10.1016/j.jad.2021.06.005

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

Karami, H. (2012). An introduction to differential item functioning. *The International Journal of Educational and Psychological Assessment, 11*(2), 59–76. https://www.researchgate.net/publication/310153908_An_Introduction_to_Differential_Item_Functioning

Keresztes, A., Ngo, C. T., Lindenberger, U., Werkle-Bergner, M., & Newcombe, N. S. (2017). Hippocampal maturation drives memory from generalization to specificity. *Trends in Cognitive Sciences, 21*(4), 249–267. https://doi.org/10.1016/j.tics.2018.05.004

Koljatic, M., Silva, M., & Sireci, S. G. (2021). College Admission Tests and Social Responsibility. *Educational Measurement: Issues and Practice, 40*(4), 22–27. https://doi.org/10.1111/emip.12425

Lucey, C. R., & Saguil, A. (2020). The consequences of structural racism on MCAT scores and medical school admissions: The past is prologue. *Academic Medicine, 95*(3), 351–356. https://doi.org/10.1097/ACM.0000000000002939

Miranda, N. A. R. (2020, January 6). Chilean university admissions tests hit by fresh protests. Reuters. https://www.reuters.com/article/us-chile-protests-university-idUSKBN1Z522I/

Moskowitz, J. B. (2022). *The hitchhiker's guide to differential item functioning (DIF)*. https://www.apadivisions.org/division-5/publications/score/2022/01/differential-item-functioning

Pedrajita, J. Q. (2015). Using Contingency Table Approaches in Differential Item Functioning Analysis: A Comparison. *Education Journal, 4*(4), 139–148. https://doi.org/10.11648/j.edu.20150404.11

Pope, D. (2023). Using Desmos and GeoGebra to engage students and develop conceptual understanding of mathematics. In C. Martin, B. Miller, & D. Polly (Eds.), *Technology Integration and Transformation in STEM Classrooms* (pp. 104–129). IGI Global Scientific Publishing. https://doi.org/10.4018/978-1-6684-5920-1.ch006

Rio, G. (2021, May 15). University of California Will No Longer Consider SAT and ACT Scores. *The New York Times*. https://www.nytimes.com/2021/05/15/us/SAT-scores-uc-university-of-california.html

Rustam, A., Naga, D., & Supriyati, Y. (2019). A comparison of Mantel-Haenszel and standardization methods: Detecting differential item functioning. *MaPan: Jurnal Matematika dan Pembelajaran, 7*(1), 16–31. https://doi.org/10.24252/mapan.2019v7n1a2

Sireci, S. G. (2021). NCME Presidential Address 2020: Valuing educational measurement. *Educational Measurement: Issues and Practice, 40*(1), 7–16. https://doi.org/10.1111/emip.12415

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292–1306. https://doi.org/10.1037/0021-9010.91.6.1292

Tan, C. Y. (2024). Socioeconomic status and student learning: insights from an umbrella review. *Educational Psychology Review, 36*(100). https://doi.org/10.1007/s10648-024-09929-3

Wetzel, E., & Böhnke, J. (2017). Differential item functioning. *Encyclopedia of Personality and Individual Differences*, 1–5. https://doi.org/10.1007/978-3-319-28099-8_1297-1

Wiberg, M. (2007). Measuring and detecting differential item functioning in criterion-referenced licensing test: A theoretic comparison of methods (EM No. 60). Umeå University, Department of Measurement and Psychometrics. https://paperzz.com/doc/7172905/measuring-and-detecting-differential-item

Wu, X., Wu, R., Peabody, M., & O'Neill, T. (2018). Detecting cross-cultural differential item functioning for increasing validity: an example from the American Board of family medicine in-training examination. *Education in The Health Professions, 1*(1), 19–23. https://doi.org/10.4103/EHP.EHP_12_18

Contact email: arlenenmendoza1@gmail.com