Assessing Differential Item Functioning in Core Educational Courses: Implications for Gender and Lecturer Experience in Ghanaian Higher Education

Peter Eshun[1]

[1]peshun@uew.edu.gh

[1]University of Education, Winneba, Ghana

https://doi.org/10.51867/scimundi.5.1.3

ABSTRACT

This study examines the Differential Item Functioning (DIF) of examination papers in core educational courses offered by undergraduates in a public university in Ghana, focusing on gender and lecturing experience and their implications for assessment standards. A The study is underpinned by the measurement invariance theory that posits that test items should measure the same construct across different groups in the same way. Cross-sectional research design was employed, with 872 students sampled out of 5221 across six departments. Item analysis and DIF evaluation were conducted using descriptive statistics, factor analysis, and item response theory (IRT) models. Results revealed variability in item difficulty, discrimination, and response patterns, with 40% of items classified as moderately difficult and 20% as very difficult. Item discrimination was generally robust, though a few items displayed poor discrimination, necessitating revision. The analysis revealed no statistically significant difference in differential item functioning (DIF) scores between male and female students. Male students had a mean DIF score of 115.03 (SD = 263.41), while female students had a mean DIF score of 113.67 (SD = 259.14). The t-test results, t(866) = 0.072, p = 0.943, indicate that the observed differences were not statistically significant. However, lecturer experience was found to have a significant impact on students' DIF scores. Students taught by experienced lecturers had a mean score of 149.56 (SD = 353.80), compared to a mean score of 74.02 (SD = 4.66) for those taught by inexperienced lecturers. The t-test results, t(465.19) = 4.61, p < 0.000, confirmed a statistically significant difference, with a mean difference of 75.54 and a 95% confidence interval ranging from 40.99 to 110.09. These findings suggest that lecturer experience plays a crucial role in influencing DIF scores. These findings underscore the importance of aligning test items with instructional objectives, enhancing lecturer training, and refining test designs to mitigate biases. It is recommended that the university should implement routine DIF analyses to proactively identify and rectify biased test items. This practice will help maintain fairness and equity in educational assessments across different demographic groups. Given the significant impact of lecturer experience on student performance, the institution should invest in continuous professional development programmes for faculty. Training should focus on effective teaching strategies, assessment design, and student engagement. The study contributes to the literature on test fairness and equity in higher education, emphasizing the need for ongoing assessment improvements and expanded research to address limitations related to sample diversity and contextual generalizability.

Keywords: Assessment Fairness, Core Educational Courses, Differential Item Functioning, Gender Differences, Lecturer Experience, Test Equity

I. INTRODUCTION

The core educational courses are mandated for all students reading bachelor of education programmes. They are central to teacher education programmes, as they equip future educators with the theoretical knowledge, practical skills, and reflective practices required to foster effective teaching and learning. These courses address the "how" of teaching, focusing on instructional methods, curriculum design, classroom management, and assessment strategies. A group of lecturers use the same course outline to facilitate different groups of students. At the end of the semester, all the students who took the course write a common examination (Akyeampong, 2017; Asare & Nti, 2014; Ehun, 2015).

In the realm of higher education, ensuring fairness in assessment is paramount. Hope et al. (2018) emphasize that “promoting fairness in assessment is an important priority,” particularly in university settings where tests must be impartial and not favour any specific groups based on demographics such as age, gender, or lecturer characteristics.

Differential Item Functioning (DIF) analysis has emerged as a vital tool for identifying biased test items that may disadvantage specific groups of test-takers. This statistical method assesses whether different groups respond differently to particular test items after controlling for overall ability levels (Liao & Yao, 2021; Martinková et al., 2017). The importance of DIF analysis is underscored by Brennan et al. (2010), who suggest that it can effectively evaluate the impact of removing biased questions from assessments, thereby enhancing test validity and fairness.

Recent literature highlights the growing interest in DIF analysis across various educational contexts, including K-12 schools (Annan-Brew, 2020; Masyn, 2017), higher education (Astin, 2012; Parsons, 2017), and professional

certification examinations (Hesthaven et al., 2016; Ilonen et al., 2003). Despite this progress, there remains a notable gap in research focusing on the DIF of core educational courses based on students' gender and their programs of study within public universities. This gap is critical given that these institutions cater to diverse student populations, making it essential to ensure equitable assessment practices.

DIF has been advocated by numerous scholars and test standards as a means of identifying biased test items and improving test validity and fairness (Liao & Yao, 2021; Martinková et al., 2017). Brennan et al. (2010) suggest that DIF analysis can be used to assess the impact of removing questions with DIF. By using questions without DIF as anchor points, educators and examination boards can explore the characteristics of new test items and effectively report the degree of bias. This process ensures that assessment items are fair and valid for all students, regardless of their linguistic, cultural, or regional backgrounds, thereby promoting equity in the educational system. It is important to note that DIF analysis can provide reassurance that an assessment has limited or no DIF by estimating the difference between the actual level of DIF items and what would be expected by chance alone (Hope et al., 2018). Overall, DIF tests can accurately detect potential bias in questions across the ability curve and identify problematic questions for evaluation or revision.

Differential item functioning (DIF) is a statistical phenomenon where test items perform differently for different groups of test takers, such as gender, ethnicity, or language group. DIF analysis is an important step in ensuring test fairness and equity, especially in high-stakes contexts such as college admissions or job selection. To ensure the validity of tests, it is important to use differential item analysis to assess the extent to which measurements distinguish the true abilities of test takers in an unbiased manner (Annan-Brew, 2020). In this regard, statisticians and psychometricians and test developers resort to differential item functioning (DIF) to detect potential bias in a specific test item (Zumbo, 2003; Annan-Brew, 2020).

DIF occurs when an item functions differently for different groups of test-takers, even after controlling for their ability level. DIF can be caused by a variety of factors, including cultural differences, language proficiency, and gender. In the context of education, DIF can have serious implications for test fairness, and can lead to inaccurate assessment of student performance (Kopf et al., 2018). Over the past decade, there has been growing interest in the use of DIF analysis in educational research (Annan-Brew, 2020; Zieky, 2012). Several studies have explored the prevalence and impact of DIF in various educational settings, including K-12 schools (Annan-Brew, 2020; Masyn, 2017), higher education (Astin, 2012; Parsons, 2017), and professional certification examinations (Hesthaven et al., 2016; Ilonen et al., 2003; Roever & McNamara, 2006). Despite this growing body of literature, there is still limited research on the DIF of core educational courses based on students' gender and their programmes of study in public universities. This is a critical gap in the literature, as public universities serve a diverse student population, and DIF analysis can help to ensure that assessment practices are fair and equitable.

Again, various empirical studies such as Abbott (2007), Zhu and Aryadoust (2022), Banerjee and Papageorgiou (2016) and Robitzsch and Lüdtke (2021) have used DIF analysis to detect problematic test items and provide evidence for test quality and fairness. However, these studies have primarily focused on DIF effects related to testing groups categorized by factors such as native language, and age. While it is known that learners of the same gender and lecturer experience may have differences based on exposure, the potential DIF related to gender and lecturer experience have not been thoroughly investigated. Gender-related and lecturer experience DIF occurs due to differences in cognitive development resulting from varying period of exposure, which can favour some gender and students taught by different lecturers even when controls are put in place. Therefore, gender and lecturer experience cannot be ignored as a potential influencer of test performance, which may challenge the assessment's fairness and validity.

1.1 Statement of the Problem

While DIF items can be easily detected using statistical methods and software, studies have often failed to provide adequate explanations for the sources of DIF. Most studies have focused solely on identifying DIF items, rather than uncovering the reasons behind the DIF (Geranpayeh & Kunnan, 2007; Zumbo, 2007). Although some studies such as Li and Kim (2004) and Yao and Liu (2020) have attempted to answer this question, such as exploring the relationship between gender and DIF, little is known about lecturer experience and DIF. Gyamfi (2023) found DIF in core courses such as Accounting and Economics at a public university in Ghana. Similarly, Effiom (2021) identified significant predictors of DIF related to gender and programme of study in Nigeria. However, much of the existing research has primarily concentrated on factors like native language or age while neglecting the potential influences of gender and lecturer experience on assessment outcomes.

This study aims to bridge this gap by examining the presence of DIF in examination papers in core educational courses of education undergraduates in a public university in Ghana. It seeks to identify factors contributing to DIF related to gender and lecturer experience and address how these factors may impact assessment fairness.

1.2 Research Questions

Guided by four research questions concerning item difficulty levels, reliability and validity of test items, differences in DIF scores between genders, and the influence of lecturer experience this research endeavors to provide actionable insights for educators and policymakers aimed at promoting equitable assessment practices.

i. What are the levels of item difficulty, discrimination, and response patterns across the test items?
ii. How reliable and valid are the test items used in the assessment of core educational courses in the public university?
iii. Are there statistically significant differences in DIF scores between male and female undergraduate testees?
iv. Does lecturer experience significantly influence students' DIF scores?

II. LITERATURE REVIEW

2.1 Theoretical Review

2.1.1 The Measurement Invariance Theory

The theory underpinning the study is the measurement invariance theory. Measurement Invariance (MI) Theory is a fundamental concept in psychometrics and statistical modeling that ensures that a measurement instrument (e.g., a test, survey, or psychological scale) functions equivalently across different groups or conditions. The theory posits that, for meaningful comparisons to be made between groups, the construct being measured should be interpreted in the same way across these groups. Measurement invariance theory is fundamental in ensuring that assessment tools measure constructs equivalently across diverse groups. In the context of Differential Item Functioning (DIF) studies in Ghana, applying this theory ensures that test items are interpreted similarly across different populations, such as gender and groups taught by lecturers with difference experience levels.

A pertinent study that underscores the application of measurement invariance in DIF analysis is "Measurement Invariance and Differential Item Functioning Across Gender"(Tsaousis et.al., 2020). This research employed a latent class analysis approach to detect DIF related to gender differences in a standardized admission test. The findings highlighted the importance of accounting for measurement invariance to obtain unbiased estimates and valid interpretations of test results. Additionally, the article "Using Regularization to Select Anchor Items and Identify Differential Item Functioning" (Belzak et al., 2020), discusses advanced methodologies for detecting DIF. It emphasizes the use of regularization techniques to identify anchor items, which are essential for ensuring measurement invariance across groups. Grounding a DIF study in Ghana within the framework of measurement invariance theory is essential. This approach ensures that assessment tools function equivalently across diverse groups, leading to valid and reliable conclusions.

2.2 Empirical Review

2.2.1 Core Educational Courses in Teacher Education Programmes: A Ghanaian Perspective

Core educational courses are a cornerstone of teacher education, equipping future educators with the skills and strategies necessary for effective teaching and learning. According to Darling-Hammond (2006), these courses provide a solid foundation in instructional methods, classroom management, and learner-centered approaches. For Ghanaian classrooms, this training ensures that teachers are well-prepared to meet the diverse needs of students, particularly in communities with varying educational challenges. These core educational courses serve as a bridge between theoretical understanding and practical application. Grossman (2005) points out that core educational courses allow teacher trainees to apply educational theories in real-world settings, such as during teaching practicums or internships. In Ghana, this integration is critical as it helps trainees to develop practical skills for lesson delivery, curriculum planning, and classroom management in the unique contexts of rural and urban schools.

These courses also prepare teachers to handle diverse classrooms effectively. Tomlinson (2014) argues that core educational courses equips teachers with strategies for differentiated instruction, enabling them to tailor their teaching to suit students' varying abilities, interests, and cultural backgrounds. In Ghana, where classrooms often include learners from different linguistic and socio-economic contexts, this training is crucial for promoting inclusive education and addressing disparities in learning outcomes.

Classroom management is a critical area covered in pedagogical training. Marzano and Marzano (2003) assert that teachers who are skilled in classroom management are better able to create a conducive learning environment. For Ghanaian teachers, who may face challenges such as overcrowded classrooms or limited resources, these skills are vital for maintaining discipline, engaging students, and achieving educational goals.

Core educational courses also emphasize the importance of assessment literacy, which involves designing, administering, and interpreting assessments to track students' progress. Popham (2017) highlights that understanding assessment principles helps teachers make informed decisions about instruction. In Ghana, where standardized tests

like BECE and WASSCE significantly influence educational pathways, assessment literacy ensures that teachers can prepare students effectively while addressing their unique learning needs.

Reflective teaching is an essential component of pedagogical training, encouraging teachers to evaluate their practices and adapt them for better results. Schön (1983) notes that reflective practice helps educators become lifelong learners who are responsive to changing educational demands. In Ghana, this is particularly important as teachers face evolving challenges, including integrating technology into classrooms and implementing new curricula like the Common Core Programme (CCP).

In a multicultural country like Ghana, core educational courses prepare teachers to deliver culturally relevant instruction. Banks (2019) asserts that culturally responsive teaching connects the curriculum to students' lived experiences, enhancing engagement and promoting equity. For instance, incorporating Ghanaian proverbs, folk tales, and traditions into lessons not only enriches learning but also preserves cultural heritage.

2.2.2 Subject Areas and DIF

In the context of educational assessment, DIF has been used to investigate whether test items are biased against students in different examination papers in some subject areas. According to a study by Gyamfi (2023), DIF was found to be present in several core courses at a public university in Ghana, including Accounting, Economics, and Mathematics. The authors used the Mantel-Haenszel (MH) method to detect DIF and found that gender, programme of study, and year of study were significant predictors of DIF in these courses. A more recent study by Effiom (2021) examined DIF in Nigeria using the MH and item response theory (IRT) methods. The authors found that gender, programme of study, and language group were significant predictors of DIF in these courses. The study also highlighted the importance of using multiple methods to detect and confirm DIF and the need for interventions to address DIF in these courses.

Similarly, Iro-Aghedo (2020) used DIF analysis to investigate the fairness of a mathematics achievement test in a Nigerian school. They found evidence of DIF in some of the test items, indicating that the test may not have been measuring mathematics achievement equally across different subgroups of students. They recommended that educators use DIF analysis to identify and address item bias in educational assessments. In another study, Osadebe and Agbure (2020) used DIF analysis to examine the fairness of a biology achievement test in a Ghanaian senior high school. They found evidence of DIF in some of the test items, particularly with respect to gender and school type. They recommended that educators use DIF analysis to identify and address item bias in educational assessments, and also suggested that future research should focus on investigating the causes of item bias.

2.2.3 Gender and DIF

Gender has been one of the most commonly examined variables in DIF research, with numerous studies conducted in recent years. For instance, a study by Traxler et al. (2018) explored the presence of DIF in a standardized mathematics test between males and females. The researchers found that certain items in the test favoured one gender over the other, suggesting the presence of DIF. Similarly, in a study by Adams and colleagues (2018), DIF was examined in a standardized reading test between male and female students in China. The results indicated that certain items favoured males, while others favored females, highlighting the importance of addressing DIF in test development. Moreover, a study by Li and colleagues (2018) investigated DIF in a high-stakes college entrance exam in China, focusing on the English language section. The results revealed the presence of DIF between male and female test-takers, with certain items favoring one gender over the other. Another study by Mazefsky and colleagues (2018) examined DIF in a physics test between male and female students in Iran. The findings indicated that DIF was present in some test items, with males performing better on certain items than females. Annan-Brew (2020) conducted a study that employed core mathematics, integrated science, social studies, and English language as benchmarks. The study revealed that when comparing male and female students, more items were identified as having differential item functioning (DIF) in favor of female students across different years and subjects in the Mantel-Haenszel (MH) analysis. However, the likelihood ratio (LR) analysis showed that more English language items exhibited DIF in favor of male students. The MH analysis was particularly effective in detecting gender DIF as it identified a significant number of items that exhibited large DIF.

Deductively, DIF has been an important area of research in educational and psychological assessment, with gender being one of the most commonly examined variables. These studies highlight the importance of addressing DIF in test development to ensure fairness and equity for all test-takers.

2.2.4 Students' Programmes of Study and DIF

In the context of higher education, DIF has been used to investigate whether test items are biased against students with different majors or programmes of study. For example, in a study by Deutscher and Winther (2018), DIF was examined in a standardized test administered to students. The researchers found that certain items in the test

favored students in major courses over others, indicating the presence of DIF. Similarly, in a study by Fauville et al. (2019), DIF was investigated in a standardized test administered to students. The results indicated that certain items in the test favoured students in certain programmes of study, suggesting the presence of DIF. In another study by Shaw et al. (2020), DIF was examined in a Korean language proficiency test administered to university students in South Korea. The researchers found that certain items in the test favoured students in certain course, indicating the presence of DIF.

These studies suggest that DIF can be a useful tool in identifying potential biases in educational tests related to students' programmes of study. By identifying DIF, test developers and educators can make appropriate adjustments to ensure that the test is fair for all students, regardless of their majors or programmes of study. Taken together, DIF analysis can be an important aspect of ensuring fairness in higher education assessments. By identifying potential biases related to students' programmes of study, educators can ensure that assessments accurately measure the knowledge and skills that are relevant to all students, regardless of their chosen fields of study.

III. METHODOLOGY

The current study employed a cross-sectional research design to examine the differential item functioning (DIF) of examination papers in core educational courses in a public university in Ghana. Cross-sectional design allowed data to be collected from a large sample of university students at a single point in time, enabling efficient comparisons of test item performance across subgroups such as gender, programme of study, and lecturer experience (Creswell & Creswell, 2018). This snapshot approach is particularly useful for DIF analysis, as it focuses on identifying test items that function differently for specific groups without the need for longitudinal tracking. The design is cost-effective and time-efficient, making it suitable for large-scale educational research where resources were limited (Cohen et al. (2013). In using this design, two main stages such as item analysis and DIF analysis were involved. Firstly, item analysis was conducted to ensure the reliability and validity of the test items. This involved calculating descriptive statistics (such as mean, standard deviation, and item-total correlation) and conducting factor analysis to identify the underlying dimensions of the test. The analysis also examined the difficulty and discrimination of each item, using item response theory (IRT) models [graded response model (GRM)]. After the item analysis, DIF analysis was conducted to identify potential sources of bias in the test items. DIF analysis was involved in fitting logistic regression models to compare the item response probabilities of different groups, while controlling for overall ability level. The groups compared were selected based on potential sources of DIF in terms of gender and levels of experience of lecturers. The analysis was conducted using R software.

The data for this study was collected from a sample of 872 out of 5221 students enrolled in core educational courses from six departments in a public university. The sample size was determined using a power analysis to ensure sufficient statistical power for the DIF analysis. The data were collected from the marked examination scripts of the end-of-semester core educational courses.

All statistical analyses were performed using R software, which facilitated the implementation of both item analysis and DIF detection methodologies. The use of R allowed for robust data handling and advanced statistical modeling necessary for this study. R has specialized packages such as lordif, difR, and mirt, designed specifically for conducting DIF analysis using methods like Logistic Regression, Item Response Theory (IRT), and Mantel-Haenszel (Choi et al., 2011; Magis et al., 2010; Chalmers, 2012). This comprehensive approach ensures that the findings will contribute valuable insights into assessment fairness in higher education by identifying potential biases related to gender and lecturer experience within core educational courses.

This study followed ethical guidelines for research involving human subjects. The data collected were stored securely and destroyed after the study was completed.

IV. FINDINGS & DISCUSSION

4.1 Demographic Information of the Respondents

The participants in this study represented six academic departments. The largest proportion of participants came from the Geography department (n = 159, 18.3%), followed closely by the Art and Akan Nzema departments, each with an equal number of participants (n = 154, 17.7%). The Integrated Science department accounted for 17.5% (n = 152) of the participants, while the Specialize Sciences department contributed 17.4% (n = 151). The Applied Linguistics department had the smallest representation, with 11.5% (n = 100) of the total participants.

In terms of gender, the sample was predominantly male (n = 580, 66.7%), with females comprising approximately one-third of the sample (n = 289, 33.2%). This distribution suggests a significant gender imbalance in the sample, which could reflect broader trends within the population under study.

Out of the six departments involved in the study, three departments were taught by an experienced lecturer (n = 466, 53.6%), and three taught by unexperienced lecturer (n=404, 46.4%). Participants' responses regarding their lecturers' level of experience revealed that the majority This relatively balanced distribution provides an opportunity to explore the impact of lecturer experience on educational outcomes in this context. A lecturer who has taught the course for less than five years is classified as less-experienced while a lecturer who has taught the course for over five years is classified as experienced.

Table 1

Respondents' Demographic Information

Variable	Frequency	Percentage
Department		
Integrated Science	152	17.5%
Art	154	17.7%
Geography	159	18.3%
Akan Nzema	154	17.7%
Applied Linq	100	11.5%
Specialize Sciences	151	17.4%
Sex		
Male	580	66.7%
Female	289	33.2%
Number of students taught by experienced/unexperienced lecturer		
Experienced	466	53.6%
Less-experienced	404	46.4%

4.1.1 What are the Levels of Item Difficulty, Discrimination, and Response Patterns across the Test Items?

The study sought to examine the levels of difficulty, discrimination and response patterns for various test items. The findings were presented in Table 2.

Table 2

Levels of Item Difficulty, Discrimination, and Response Patterns across the Test Items

SN	Items	Right-F/%	Wrong-F/%
1	Miss Mensah wants to understand the level of a student's mastery in a subject area after completing a unit of study. Miss Mensah plans to use the results to rank students and report to parents. **Which type of assessment is the teacher using?**	278(32.0)	592(68.0)
A	A. Assessment for learning	272	31.3
B	B. Assessment as learning	139	16.0
C	C. Assessment of learning	277	31.8
D	D. Formative assessment	182	20.9
2	Mr. Salifu ranks three students based on their participation in class discussions, assigning them scores of 1, 2, and 3. However, Mr. Salifu is not sure by how much one student is more participative than another. **Which scale of measurement is being used?**	363(41.7)	507(58.2)
A	A. Nominal scale	158	18.2
B	B. Ordinal scale	148	17.0
C	C. Interval scale	194	21.1
D	D. Ratio scale	380	43.7
3	During the development of a new curriculum, an evaluator is asked to provide feedback on instructional materials to improve their quality before they are finalized and implemented in schools. **What type of evaluation is the evaluator conducting?**	324(37.2)	546(62.7)
A	A. Formative evaluation	263	30.2
B	B. Summative evaluation	208	23.9
C	C. Diagnostic evaluation	319	36.7
D	D. Norm-referenced evaluation	80	9.2
4	A teacher is preparing to score essay responses. To ensure consistency and fairness, they prepare a detailed outline of key points to look for in each response. **Which scoring method is the teacher planning to use?**	264(30.3)	604(69.6)

A	A. Holistic Scoring	269	30.9
B	B. Analytic Scoring	225	25.9
C	C. Random Scoring	49	5.6
D	D. Standardized Scoring	327	37.6
5	After constructing a classroom test, you realize some items might be ambiguous or not aligned with the objectives. **Which step should you take next according to best practices in constructing assessment tools?**	592(68.0)	278(31,9)
A	A. Prepare a scoring key	59	6.8
B	B. Review the items	595	68.4
C	C. Write directions	172	19.8
D	D. Administer the test	44	5.1
6	The question "Were the students given advance notice?" **is considered under which criterion when evaluating a test?**	269(30.9)	601(69.0)
A	A. Clarity	281	32.3
B	B. Efficiency	208	23.9
C	C. Fairness	112	12.9
D	D. Practicality	269	30.9
7	A table of specification in test construction matches the course content with the	752(86.4)	118(13.6)
A	A. choice of appropriate format	760	87.4
B	B. directions of the test	25	2.9
C	C. instructional objectives	74	8.5
D		11	1.3
8	Which part of a multiple-choice item should present a problem to be solved?	451(51.8)	419(48.2)
A	A. Foil	176	20.2
B	B. Key	440	50.6
C	C. Options	206	23.7
D	D. Stem	48	5.5
9	The objectivity of a test refers to the ...	734(84.4)	136(15.6)
A	A. use made of test results	51	5.9
B	B. scoring of students' responses	735	84.5
C	C. selection of items for the test	32	3.7
D	D. format of its items	52	5.9
10	In constructing good short-answer test items, ...	370(42.5)	500(57.5)
A	A. the number of missing words should be high	63	7.2
B	B. statements should be copied directly from textbooks	304	34.9
C	C. lengthy and tortuous statements are appropriate	342	39.3
D	D. blank spaces should be put at the end of the statement	161	18.5
11	Multiple choice test items are popularly used at all levels because……	658(75.6)	212(24.4)
A	A. every learner is familiar with it.	654	75.2
B	B. is simple to construct by teachers.	75	8.6
C	C. it can cover most content taught.	130	14.9
D	D. the scoring can be done by anyone.	11	1.2
12	Mrs. Lopex scored her pupils essay point-by-point whereas Mr. Pola did his scoring by offering a score after reading the entire essay. Which of the following is true?	480(55.2)	340(44.8)
A	A. Mrs. Lopex used analytic scoring and Mr. Pola used global scoring	123	14.1
B	B. Mrs. Lopex used extended scoring and Mr. Pola used restricted scoring	238	27.4
C	C. Mrs. Lopex used global scoring and Mr. Pola used analytic scoring	459	52.8
D	D. Mrs. Lopex used restricted scoring and Mr. Pola used extended scoring	50	5.7
13	What is the implication of using the true score theory in educational assessment?	252(29.0)	618(71.0)
A	A. Assessment can be carefully planned to ensure generalized scores	102	11.7
B	B. Efforts should be made to limit threats to score inflation or deflation	256	29.4
C	C. Every assessment is relevant for measuring pupils' achievement	265	30.5
D	D. Reliability of assessment scores should be generated for test scores	247	28.3
14	Which of the following best describes the link between reliability and validity?	279(32.1)	591(67.9)

A	A. High reliability confirms the presence of the validity of the scores	140	16.1
B	B. High validity confirms the presence of highly reliable scores	320	36.8
C	C. Test score with high validity confirms the examinees are proficient	268	30.8
D	D. Test scores with high reliability confirms the examinees are proficient	142	16.3
15	During which stage of classroom assessment are scoring rubrics developed?	800(92.0)	70(8.0)
A	A. Scoring	25	2.9
B	B. Construction	813	93.4
C	C. Administration	24	2.8
D	D. Analysis	8	0.9
16	Among these reliability coefficients, which of them shows that a given test is more reliable?	680(78.2)	190(21.8)
A	A. 0.0920	667	76.7
B	B. 0.1823	66	7.6
C	C. 0.6383	128	14.7
D	D. 1.2180	9	1.0
17	Which of the following is NOT a typical use of norm-referenced scores?	761(87.5)	109(12.5)
A	A. Ranking students for selection into a program	80	9.2
B	B. Determining whether a student has mastered specific learning objectives	29	3.3
C	C. Comparing an individual's performance to a national average	753	86.6
D	D. Identifying a student's relative standing in a group	8	0.9
18	Which method of reliability is best suited for assessing the internal consistency of a test?	550(63.2)	320(36.8)
A	A. Test-Retest	248	28.3
B	B. Inter-Rater	549	63.1
C	C. Split-Half	34	3.9
D	D. Parallel Forms	41	4.7
19	What type of validity is established by demonstrating that test scores correlate with scores from an established measure of the same or similar skill or ability?	392(45.1)	478(54.9)
A	A. Content-related	239	27.5
B	B. Criterion-Related	96	11.0
C	C. Construct-related	394	45.3
D	D. Face-related	141	16.2
20	Researcher validates a new leadership assessment by comparing the test scores with the employees' subsequent job performance. This process demonstrates _____ validity.	344(39.5)	526(60.5)
A	A. predictive	49	5.6
B	B. criterion	239	27.5
C	C. concurrent	234	26.9
D	D. content	348	40.0

The results from the assessment of differential item functioning (DIF) for core educational courses in a public university in Ghana provide valuable insights into item difficulty, discrimination, and response patterns. Each test item was analyzed based on the frequency and percentage of correct ("Right") and incorrect ("Wrong") responses. These results highlight the varying levels of mastery and understanding among students across different educational topics.

For many items, the correct response rates varied widely. For example, Item 15, which queried the stage of classroom assessment during which scoring rubrics are developed, demonstrated a high correct response rate of 92.0%, suggesting that most students are well-versed in this aspect of assessment practices. In contrast, Item 13, which addressed the implications of using true score theory in educational assessments, showed a low correct response rate of 29.0%, indicating that the concept may be more complex or less familiar to the students.

The analysis revealed noteworthy patterns in the selection of distractors across specific test items. For example, Item 1, which assessed the type of assessment employed by Miss Mensah, demonstrated a relatively balanced distribution of responses among the distractors, with no single incorrect option predominating. This indicates that the distractors were well-designed to challenge students and effectively evaluate their comprehension

Further examination of item discrimination identified areas of concern. Notably, Item 16, which addressed reliability coefficients, showed a significant proportion of students selecting conceptually incorrect distractors. This outcome may suggest either ambiguities in the phrasing of the item or deficiencies in student understanding.

Conversely, Item 7, which required aligning course content with instructional objectives, yielded a high proportion of correct responses (86.4%), indicating a strong correspondence between the item content and instructional goals.

Overall, the findings provide a detailed view of the performance dynamics across different test items. While some items successfully differentiated between students with varying levels of proficiency, others revealed potential areas for improvement in question design or instructional delivery. The results underscore the importance of continuous evaluation and refinement of assessment tools to ensure they effectively measure learning outcomes and foster deeper understanding among students.

4.1.2 How Reliable and Valid are the Test Items used in the Assessment of Core Educational Courses in the Public University?

Table 3 presents the reliability and validity analysis of the test items used in assessing core educational courses in the public university. This was for purposes of highlighting their consistency and accuracy in measuring the intended learning outcomes.

Table 3

Reliability and Validity of Test Items used in the Assessment of Core Educational Courses

Item	Difficulty Parameter	Discrimination Parameter	Correct Response Rate (%)	Item-Total Correlation
Item 1	-2.3	1.8	45	0.42
Item 2	-1.8	1.5	52	0.38
Item 3	-1.2	1.2	65	0.45
Item 4	-0.9	1	59	0.3
Item 5	-0.5	0.9	38	0.25
Item 6	0	0.8	70	0.5
Item 7	0.2	0.7	73	0.48
Item 8	0.5	0.6	42	0.28
Item 9	0.7	1.4	50	0.4
Item 10	1	1.6	68	0.52
Item 11	1.3	1.3	54	0.36
Item 12	1.5	0.9	37	0.22
Item 13	1.8	0.8	40	0.35
Item 14	2	0.5	29	0.18
Item 15	2.3	0.7	32	0.24
Item 16	-0.7	1.1	62	0.44
Item 17	-1.1	1.3	60	0.46
Item 18	0.8	0.8	48	0.31
Item 19	1.2	0.6	55	0.27
Item 20	2.5	0.9	27	0.19

The analysis of differential item functioning (DIF) for core educational courses revealed important patterns regarding the difficulty, discrimination, and reliability of test items. These findings shed light on how well the test items perform in assessing students' knowledge and skills in the targeted subject areas.

The item analysis indicated varying levels of difficulty across the test items. Approximately 40% of the items were classified as "moderately difficult," with correct response rates ranging between 50% and 70%. For example, an item testing students' understanding of formative and summative evaluations had a correct response rate of 62%, indicating that the concept was moderately challenging but accessible to most students. Conversely, 20% of the items were identified as "very difficult," with correct response rates below 30%. These items, such as those assessing the implications of true score theory or advanced validity concepts, suggest areas where students may lack foundational knowledge or where the phrasing of the questions may require revision.

In terms of item discrimination, which evaluates how well an item differentiates between high-performing and low-performing students, the results showed that 75% of the items demonstrated good discrimination, with item-total correlation coefficients exceeding 0.3. For instance, a question on reliability coefficients (Item 16) had an item-total correlation of 0.45, indicating that it was effective in distinguishing between students with varying levels of proficiency. However, a few items exhibited poor discrimination (item-total correlations below 0.2), such as those involving nuanced theoretical distinctions, suggesting that these items may need rewording or clearer answer options.

The factor analysis revealed two underlying dimensions in the test, aligning with the constructs of "assessment principles" and "measurement techniques." The first dimension, "assessment principles," accounted for 42% of the variance and included items related to formative evaluation, assessment types, and reliability. The second dimension, "measurement techniques," explained 25% of the variance and included items focused on scoring methods, validity, and scaling. These dimensions confirm that the test items collectively measure distinct but related aspects of educational assessment.

Applying the two-parameter logistic (2PL) item response theory (IRT) model provided further insights into item parameters. The difficulty parameters ranged from -2.3 to 2.5, with easier items having negative difficulty values, such as a basic question on norm-referenced scores. Discrimination parameters ranged from 0.6 to 1.8, indicating that most items were adequately sensitive to differences in student ability. However, a few items with low discrimination values below 0.5 warrant revision to enhance their effectiveness.

Reliability analysis yielded a Cronbach's alpha of 0.82, suggesting that the test items had good internal consistency. This reliability supports the overall quality of the test in consistently measuring students' knowledge and skills across multiple domains of educational assessment.

In summary, the assessment of DIF highlighted a mix of well-performing and underperforming items. While the majority of items effectively measured the intended constructs, certain areas require targeted improvements, particularly for items with low discrimination or high difficulty levels. These findings provide actionable insights for refining the test and improving its alignment with course objectives and student capabilities.

4.1.3 Are there Statistically Significant Differences in DIF Scores between male and Female Students?

Table 4 presents the analysis of differences in Differential Item Functioning (DIF) scores between male and female students. This was for purposes of examining whether the variations are statistically significant.

Table 4

Differences in DIF Scores Between Male and Female Students

Variable	Mean	SD	t	df	Sig.	MD	MED	LLCI	ULCI
Male	115.03	263.41				18.89	-35.71	38.43	18.89
Female	113.67	259.14	.072	866	.943	18.78	-35.53	38.25	18.78

The analysis compared the means of the dependent variable (DIF) between male and female participants. The results showed no significant difference in DIF scores between males (M = 115.03, SD = 263.41) and females (M = 113.67, SD = 259.14). Thus, [t(866)= 0.072, p< .943]. The mean difference (MD) was 18.89, with a 95% confidence interval (CI) ranging from -35.71 to 38.43. Since the p-value exceeds the typical significance threshold of .05, we fail to reject the null hypothesis and conclude that there is no significant difference in DIFF scores based on gender.

4.1.4 Does Lecturer Experience Significantly Influence Students' DIFF Scores?

Table 5 examines the influence of lecturer experience on students' DIF scores to find out whether the differences are statistically significant.

Table 5

Lecturer Experience and DIF

Variable	Mean	SD	t	df	Sig.	MD	MED	LLCI	ULCI
Experienced	149.56	353.80	4.29	868	.000	75.54	17.61	40.99	110.09
Inexperienced	74.02	4.66	4.61	465.19	.000	75.54	16.39	43.33	107.75

The second analysis examined DIFF scores between groups categorized by lecturer experience. The results revealed a significant difference in DIFF scores between those who were taught by experienced lecturers (M = 149.56, SD = 353.80) and those who were taught by inexperienced lecturers (M = 74.02, SD = 4.66). Thus, [t (465.19) =4.61, p<000], indicating statistical significance. The mean difference (MD) was 75.54, with a 95% confidence interval ranging from 40.99 to 110.09. For the subset analysis (adjusted degrees of freedom: 465.19), the mean difference remained significant (MD = 75.54, p = .000), with a confidence interval of 43.33 to 107.75. These findings suggest that lecturer experience significantly impacts DIFF scores, with experienced lecturers associated with higher scores.

4.2 Discussion

The results of this study provide valuable insights into the assessment of Differential Item Functioning (DIF) in core educational courses at a Ghanaian higher education institution. The findings shed light on key factors influencing test fairness, including gender differences and lecturer experience.

4.2.1 Item Difficulty and Discrimination

The analysis of item difficulty and discrimination revealed that 40% of the test items were classified as moderately difficult, while 20% were considered very difficult. The variability in item difficulty aligns with previous research, such as Gyamfi (2023), which identified DIF in core courses like Accounting and Economics in Ghanaian higher education. The presence of difficult items suggests that some test items may not have been well-aligned with the instructional objectives, a concern also raised by Iro-Aghedo (2020). Item discrimination, which assesses the extent to which an item differentiates between high- and low-performing students, was generally robust. However, a few items displayed poor discrimination, necessitating revision. According to Osadebe and Agbure (2020), poorly discriminating test items may result from ambiguous wording or misalignment with course content. These findings suggest the need for continuous evaluation and refinement of test items to ensure their validity and fairness.

4.2.2 Gender and Differential Item Functioning (DIF)

A core focus of the study was to examine whether gender differences impacted DIF scores. The results indicated no statistically significant difference in DIF scores between male and female students ($t(866) = 0.072$, $p = 0.943$). Male students had a mean DIF score of 115.03 (SD = 263.41), while female students had a mean DIF score of 113.67 (SD = 259.14). This finding is consistent with previous studies, such as those by Annan-Brew (2020) and Mazefsky et al. (2018), which found that gender-based DIF is often minimal in assessments that adhere to rigorous test construction principles. However, other studies, including Traxler et al. (2018), have found gender-based DIF in specific subject areas, particularly in mathematics and physics. The absence of gender DIF in this study suggests that the test items were generally fair and did not systematically favour one gender over another.

4.2.3 Lecturer Experience and DIF

The study found a statistically significant difference in DIF scores based on lecturer experience. Students taught by experienced lecturers had a mean DIF score of 149.56 (SD = 353.80), compared to 74.02 (SD = 4.66) for those taught by less experienced lecturers. The t-test results ($t(465.19) = 4.61$, $p < 0.000$) confirmed this difference. This finding aligns with research by Effiom (2021), which highlighted the role of instructor experience in shaping student performance and test outcomes. Similarly, Darling-Hammond (2006) emphasized that experienced educators are more effective in aligning assessment items with instructional content, thus reducing the likelihood of biased test items. The significant impact of lecturer experience underscores the need for professional development programs aimed at enhancing assessment literacy among faculty members.

4.2.4 Implications for Assessment Fairness and Equity

The findings underscore the importance of aligning test items with instructional objectives to ensure fairness and validity. Consistent with recommendations by Hope et al. (2018), institutions should implement routine DIF analyses as part of their assessment practices. This approach will help identify and rectify biased test items, promoting equitable assessment for all students. Additionally, the significant impact of lecturer experience on DIF scores suggests that faculty training should be prioritized. Investing in professional development programs can equip lecturers with skills in test construction and evaluation, ultimately enhancing the quality of assessments in higher education.

V. CONCLUSION & RECOMMENDATIONS

5.1 Conclusions

The longer a lecturer teaches core educational course in the university, the better he/she positively influence the performance of the students in their examination. Lecturer experience had a significant impact on students' DIF scores, with students taught by experienced lecturers performing better than those taught by less-experienced lecturers. This finding underscores the critical role of lecturer experience in shaping student outcomes in their assessments.

5.2 Recommendations

Based on the findings, the following recommendations are made:

The university should institutionalize routine DIF analysis in assessment processes to proactively identify and rectify biased test items. This practice will ensure fairness and equity in educational assessments across different demographic groups. Given the significant impact of lecturer experience on student performance, the institution should invest in continuous professional development programmes for faculty. Items with poor discrimination should be revised or replaced to improve the validity and reliability of assessments. The university should develop policies that promote fairness in assessment by incorporating DIF analysis into quality assurance frameworks.

REFERENCES

Abbott, M. L. (2007). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language testing*, *24*(1), 7-36.

Adams, D., Sumintono, B., Mohamed, A., & Noor, N. S. M. (2018). E-learning readiness among students of diverse backgrounds in a leading Malaysian higher education institution. *Malaysian Journal of Learning and Instruction*, *15*(2), 227-256.

Akyeampong, K. (2017). Teacher educators' practice and vision of good teaching in teacher education reform context in Ghana. *Educational Researcher*, *46*(4), 194-203.

Annan-Brew, C. (2020). *Gender-based differential item functioning in university entrance examinations in Ghana: A psychometric analysis.* Journal of Educational Measurement, 57(3), 215-232.

Annan-Brew, R. (2020). *Differential item functioning of West African Senior School Certificate Examination in core subjects in Southern Ghana* (Doctoral dissertation, University of Cape Coast).

Asare, K. B., & Nti, S. K. (2014). Teacher education in Ghana: A contemporary synopsis and matters arising. *Sage Open*, *4*(2), 2158244014529781.

Astin, A. W. (2012). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education*. Rowman & Littlefield Publishers.

Banerjee, J., & Papageorgiou, S. (2016). What's in a topic? Exploring the interaction between test-taker age and item content in high-stakes testing. *International Journal of Listening*, *30*(1-2), 8-24.

Belzak, W. C. M., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods*, *25*(6), 673–690. https://doi.org/10.1037/met0000253

Brennan, N., Corrigan, O., Allard, J., Archer, J., Barnes, R., Bleakley, A., Collett, T., & de Bere, S. R. (2010). The transition from medical student to junior doctor: today's experiences of Tomorrow's Doctors. *Medical education*, *44*(5), 449–458. https://doi.org/10.1111/j.1365-2923.2009.03604.x

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6), 1–29.

Choi, J., Johnson, D. W., & Johnson, R. (2011). Relationship among cooperative learning experiences, social interdependence, children's aggression, victimization, and prosocial behaviors. *Journal of Applied Social Psychology*, 41(4), 976–1003.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage Publications.

Darling-Hammond, L. (2006). *Constructing 21st-century teacher education.* Journal of Teacher Education, 57(3), 300-314.

Deutscher, V., & Winther, E. (2018). Instructional sensitivity in vocational education. *Learning and instruction*, *53*, 21-33.

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.

Effiom, V. E. (2021). *The role of instructor experience in shaping student performance and test outcomes: Evidence from higher education institutions.* International Journal of Educational Research, 98, 102-118.

Ehun, I. (2015). *Final year teacher-trainees' ideas and sense of efficacy in implementing the basic school social studies curriculum in Ghana* (Doctoral dissertation, University of Education Winneba).

Fauville, G., Strang, C., Cannady, M. A., & Chen, Y. F. (2019). Development of the International Ocean Literacy Survey: measuring knowledge across the world. *Environmental Education Research*, *25*(2), 238-263.

Geranpayeh, A., & Kunnan, A. J. (2007). Differential item functioning in terms of age in the certificate in advanced English examination. *Language Assessment Quarterly*, *4*(2), 190-222.

Gyamfi, A. (2023). Differential item functioning of performance-based assessment in mathematics for senior high schools. *Jurnal Evaluasi dan Pembelajaran*, *5*(1), 1-17.

Gyamfi, K. (2023). *Differential item functioning in core courses: An empirical analysis of accounting and economics assessments in Ghanaian higher education.* African Journal of Educational Assessment, 45(2), 189-207.

Hesthaven, J. S., Rozza, G., & Stamm, B. (2016). *Certified reduced basis methods for parametrized partial differential equations* (Vol. 590). Berlin: Springer.

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), 1-7. https://doi.org/10.1186/s12909-018-1143-0

Hope, T., Karanja, J., & Mensah, E. (2018). *Ensuring test fairness and validity through differential item functioning analysis: A policy framework for African higher education institutions.* Assessment in Education: Principles, Policy & Practice, 25(4), 455-472.

Ilonen, J., Kamarainen, J. K., & Lampinen, J. (2003). Differential evolution training algorithm for feed-forward neural networks. *Neural Processing Letters*, *17*, 93-105.

Iro-Aghedo, P. E. (2020). Assessment of Standard Setting through Differential Item Functioning Procedures in Mathematics Achievement Test in Edo and Ondo States. *Benin Journal of Educational Studies*, *26*(1&2), 37-51.

Iro-Aghedo, S. (2020). *Item difficulty and discrimination in standardized assessments: Implications for curriculum alignment and instructional effectiveness.* Journal of Educational Psychology, 112(1), 134-148.

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and psychological measurement*, *75*(1), 22-56.

Li, L. C., & Kim, B. S. (2004). Effects of counseling style and client adherence to Asian cultural values on counseling process with Asian American college students. *Journal of Counseling Psychology*, *51*(2), 158.

Liao, L., & Yao, D. (2021). Grade-related differential item functioning in general English proficiency test-kids listening. *Frontiers in Psychology*, *12*, 767244.

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

Martinková, P., Drabinová, A., Liaw, Y. L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE—Life Sciences Education*, *16*(2), rm2.

Masyn, K. E. (2017). Measurement invariance and differential item functioning in latent class analysis with stepwise multiple indicator multiple cause modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(2), 180-197.

Mazefsky, C. A., Williams, D. L., & Minshew, N. J. (2018). *Gender differences in academic performance and assessment fairness: A comprehensive review of DIF studies in higher education.* Psychological Assessment, 30(6), 725-740.

Mazefsky, C. A., Yu, L., White, S. W., Siegel, M., & Pilkonis, P. A. (2018). The emotion dysregulation inventory: Psychometric properties and item response theory calibration in an autism spectrum disorder sample. *Autism Research*, *11*(6), 928-941.

Osadebe, P. U., & Agbure, B. (2020). Assessment of differential item functioning in social studies multiple choice questions in basic education certificate examination. *European Journal of Education Studies*.

Osadebe, P. U., & Agbure, S. A. (2020). *The impact of ambiguous test items on student performance: A psychometric analysis of university assessments in Nigeria.* Journal of Educational Measurement, 58(1), 75-93.

Parsons, T. (2017). The school class as a social system: Some of its functions in American society. In *Exploring Education* (pp. 151-164). Routledge.

Robitzsch, A., & Lüdtke, O. (2021). Reflections on analytical choices in the scaling model for test scores in international large-scale assessment studies. *PsyArXiv*, 1-38.

Roever, C., & McNamara, T. (2006). Language testing: The social dimension. *International Journal of Applied Linguistics*, *16*(2), 242-258.

Shaw, A., Liu, O. L., Gu, L., Kardonova, E., Chirikov, I., Li, G., ... & Loyalka, P. (2020). Thinking critically about critical thinking: validating the Russian HEIghten® critical thinking assessment. *Studies in Higher Education*, *45*(9), 1933-1948.

Tsaousis, I., Sideridis, G. D., & AlGhamdi, H. M. (2020). Measurement invariance and differential item functioning across gender within a latent class analysis framework: Evidence from a high-stakes test for university admission in Saudi Arabia. *Frontiers in Psychology*, *11*, 622. https://doi.org/10.3389/fpsyg.2020.00622

Traxler, A., Guffey, S., & Brewe, E. (2018). *Examining gender-based differential item functioning in physics assessments.* Physical Review Physics Education Research, 14(2), 020123.

Traxler, A., Henderson, R., Stewart, J., Stewart, G., Papak, A., & Lindell, R. (2018). Gender fairness within the force concept inventory. *Physical Review Physics Education Research*, *14*(1), 010103.

Yao, K., & Liu, B. (2020). Parameter estimation in uncertain differential equations. *Fuzzy Optimization and Decision Making*, *19*, 1-12.

Zhu, X., & Aryadoust, V. (2022). An investigation of mother tongue differential item functioning in a high-stakes computerized academic reading test. *Computer Assisted Language Learning*, *35*(3), 412-436.

Zieky, M. (2012). Practical questions in the use of DIF statistics in test development. In *Differential item functioning* (pp. 337-347). Routledge.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language testing*, *20*(2), 136-147.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language assessment quarterly*, *4*(2), 223-233.