Journal Transformation Of Knowledge (JTK)
ISSN 2987-2375 (media online)
https://ojs.yptransteknologi.sch.id/index.php/jtk
Vol. 02, No. 03, November 2024

MATHEMATICS ASSESSMENT: CAUSES AND SOLUTIONS OF GENDER BIASED ITEMS

[1]Chinelo Blessing Oribhabor, University of Africa, Nigeria.
E-mail: chinelo.oribhabor@uat.edu.ng

ARTICLE INFO
Review Article
Received: 02, 10.2024.
Revised: 18, 10.2024.
Accepted: 18, 11.2024.

Keywords:
Mathematics, Assessment, Gender, Item Bias, Causes

ABSTRACT
This paper examines the phenomenon of gender-biased items in Mathematics assessments, exploring both the causes and potential solutions. Gender bias in Mathematics tests can result from various factors, including item format, context of Mathematics test item, cultural stereotypes, differential item functioning, and teacher expectations. To address these biases, several solutions are proposed. These include the use of differential item functioning analysis to detect and eliminate biased items, incorporating diverse item formats, training for educators to recognize and counteract their own biases, and designing assessments that are culturally and contextually inclusive. Additionally, promoting a growth mindset and encouraging female participation in Mathematics from an early age can help mitigate the long-term impacts of gender bias in assessments. By addressing gender bias in Mathematics assessment, the study contributes to creating a more equitable learning environment for all students.

INTRODUCTION

Mathematics is a subject that every student at the primary and secondary school levels is expected to offer. Its importance made the Federal Government of Nigeria to make Mathematics a core subject at both primary and secondary education levels (Federal Republic of Nigeria, 2013). In spite of the important role Mathematics plays in everyday life, it has remained one of the subjects students find difficult to pass in Nigerian schools (Alade, Aletan & Sokenu, 2020). According to House & Telese, (2008), various researches had been undertaken to find ways of improving Mathematics achievement and determine the factors influencing Mathematics' learning and performance. The identified factors include among others, motivational orientation, self-esteem/self-efficacy, lack of adequate preparation, shortage of qualified teachers, lack of good school environment and infrastructural facilities (Aremu and Sokan 2003), students' poor attitude towards Mathematics (Bolaji, 2005) and poor teaching methods adopted by teachers (National Mathematics Centre, NMC, 2009).

To improve performances, in Mathematics, many interventions have been initiated. Prominent among the interventions are the Lagos Eko Secondary Education project, 2004-2017 and the NMC's Mathematics Improvement Programme (MIP) aimed at creating new teaching methodologies to improve students' performance in Mathematics. Despite the interventions, the observed gradual performance persisted as evident in the fluctuating result of candidates in WASSCE's Mathematics after the introduction of these interventions (This Day Newspaper, 2024). One of the areas of challenge may be the observation that examiners are often faced with challenges of how to assess students in ways to obtain fair scores by reducing item difficulty especially in Mathematics (Olonode, 2016). As shown by Rover (2005), a fair and equitable test is one, which allows all the testees equal opportunity to exhibit the aptitudes and information which they have obtained and which are applicable to the motivation behind the test. The matter of test fairness also brings forth the issue of item bias.

The APA/NCME/AERA Standards for educational and psychological testing list fairness as an important consideration for any test, as important as validity and reliability (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014). Fairness is also intrinsically desirable: both teachers and students would like to be using and taking assessments that they believe to be fair. In addition, it is a necessary precondition for a strong validity argument (Kane, 2013). Item bias refers to the presence of unfairness or discrimination in test items, where certain subgroups of the population perform differently on the items due to factors such as gender or race and ethnicity. According to Davidson Wortzman, Ko, and Li (2021), item bias refers to the differential performance of different groups of test-takers on specific exam items. For example, if females were less likely to endorse an item from an

achievement test of mathematical ability than males because the question required prior knowledge of sports terms that the females are not familiar with, then the item is biased.

Research indicates that the format and content of test items can disproportionately affect male and female students, leading to discrepancies in performance that do not accurately reflect their mathematical abilities. The implications of such biases are profound, as they can influence educational outcomes, self-esteem, and future opportunities for students. In the context of mathematics assessments, it has been observed that certain item formats, such as multiple-choice versus constructed-response items, can favor one gender over another. For instance, studies such as Pedrajita (2009), Shear (2023) have shown that male students tend to perform better on multiple-choice items, while female students often excel in constructed-response formats. This discrepancy raises concerns about the fairness and validity of assessments used for high-stakes decisions, such as course placements and graduation requirements.

Gender bias in mathematics assessments refers to the presence of test items that systematically disadvantage one gender group over another, despite equal underlying ability. These biased items measure something other than the intended mathematical construct, leading to unfair comparisons of performance between male and female students. The persistent issue of gender bias in Mathematics assessments necessitates a thorough investigation into the specific causes and potential solutions. This paper identified the characteristics of test items that contribute to gender bias, and explored effective strategies to mitigate this bias in future assessments. Understanding these dynamics is crucial for developing fairer assessment practices that accurately reflect the capabilities of all students, regardless of gender.

Concept of Mathematics Assessment

Mathematics is being viewed not only as a traditional prerequisite subject for prospective scientists, engineers, businessman etc, but, also as a fundamental aspect of literacy. Assessment is an integral part of teaching-learning process as it is a prime tool for monitoring the progress and shaping learning. (National Council of Teachers of Mathematics, 1989). Therefore, Mathematics assessment can be defined as the systematic process of gathering, interpreting, and using information about students' mathematical learning. This process involves various methods, including formative assessments (ongoing assessments during the learning process) and summative assessments (evaluations at the end of an instructional unit). The primary purpose of these assessments is not only to evaluate student performance but also to inform instructional practices and improve learning outcomes.

Types of Mathematics Assessment

1. Formative Assessment: This type of assessment is conducted during the learning process. It aims to provide immediate feedback to students and teachers, allowing for adjustments in teaching strategies and learning activities. Techniques may include quizzes, observations, and student reflections (National Academies, 1993).
2. Summative Assessment: Summative assessments occur at the end of an instructional period and are typically used to evaluate student learning against a standard or benchmark. Examples include final exams, standardized tests, and end-of-term projects (Suurtamm, 2016).
3. Diagnostic Assessment: This type assesses students' existing knowledge and skills before instruction begins. It helps identify areas of strength and weakness, allowing educators to tailor their teaching to meet students' needs (Suurtamm, 2016).
4. Performance Assessment: This involves tasks that require students to apply their mathematical knowledge in real-world scenarios. It emphasizes problem-solving, reasoning, and the application of concepts rather than rote memorization (eGyanKosh, 2018).

Principles of Effective Mathematics Assessment

Effective mathematics assessment is grounded in several key principles:

1. Alignment with Learning Goals: Assessments should align with the curriculum and learning objectives, ensuring that they measure what students are expected to learn.
2. Equity and Accessibility: Assessments must be designed to be fair and accessible to all students, regardless of their background or learning needs. This includes considering cultural relevance and the potential for bias in assessment items.

3. Use of Multiple Measures: Relying on a variety of assessment methods provides a more comprehensive view of a student's mathematical understanding and skills. This approach helps to mitigate the limitations of any single assessment method.
4. Feedback for Improvement: High-quality assessments should provide constructive feedback that students can use to improve their understanding and performance in mathematics. This feedback loop is essential for fostering a growth mindset among learners

The design and implementation of mathematics assessments can significantly influence educational outcomes. Effective assessments can enhance student learning by identifying areas for improvement and guiding instructional practices. Conversely, poorly designed assessments can lead to misconceptions, reinforce stereotypes, and create barriers to learning.

Causes of Gender Bias in Mathematics Items

1. Item Format: The format of test items can introduce gender bias. Studies have found that multiple-choice items tend to favor male students, while constructed-response items generally favor females (Shanmugam & Kanageswari, 2020). This suggests that the skills required to answer different item formats may be differentially associated with gender.
2. Language Load: The language complexity of test items can disproportionately affect the performance of certain gender groups. Abedi and Lord (2001) argued that reading comprehension should not be considered a relevant construct in mathematics assessments, as it may introduce bias against students who excel in mathematical reasoning but struggle with language (Shanmugam & Kanageswari, 2020).
3. Content and Context: The content and context of mathematics items can also contribute to gender bias. For example, geometry, probability, and algebra items have been found to favor males, while statistical interpretation and multistep problem-solving items tend to favor females (Reardon et al., 2018).
4. Higher-Order Thinking Skills: The inclusion of higher-order thinking (HOT) items in assessments may help reduce gender bias by providing a more equitable platform for evaluating student understanding. However, research suggests that HOT computation items with minimal language load do not necessarily favor girls (Shanmugam & Kanageswari, 2020).
5. Parental and Teacher Bias: Gender-biased assessments from parents and teachers contribute to the development of self-perceptions in students. If teachers or parents hold stereotypes about gender capabilities in mathematics, this can affect their evaluations and encouragement, perpetuating the cycle of bias (Adamecz, Jerrim, Pingault & Shure, 2023).
6. Stereotype Threat: The phenomenon of stereotype threat, where individuals perform worse when they are aware of negative stereotypes about their group, can significantly impact female students in mathematics. When test administrators signal expectations of gender differences, it can lead to poorer performance among girls (Niederle &Vesterlund, 2010).

Impact of Gender Biased Items

According to (Shanmugam & Kanageswari, 2020), gender-biased items in mathematics assessments can have far-reaching consequences for students' academic and career opportunities:

1. Course Placement: If assessments used for course placement, such as tracking into advanced math classes, are biased against female students, it can limit their access to rigorous coursework and opportunities to develop their mathematical skills.
2. University Admissions: High-stakes competitive tests used for university admissions have been shown to increase gender gaps in math performance, especially for high-achieving female students.
3. Self-Confidence and Interest: Experiencing gender bias in math assessments throughout their education can negatively impact female students' self-confidence, interest, and persistence in STEM fields.
4. Labor Market Outcomes: The underrepresentation of women in STEM fields, partly driven by gender gaps in math achievement, contributes to gender segregation in the labor market, leading to disparities in earnings and career advancement opportunities.

Addressing Gender Bias in Mathematics Assessments

To mitigate gender bias in mathematics assessments, researchers and educators recommend:

1. Adopting Differential Item Functioning (DIF) analysis: DIF techniques can identify items that function differently for male and female students, allowing for the removal or revision of biased items (Alade, Aletan & Sokenu, 2020).
2. Incorporating diverse item formats: Using a variety of item formats, including multiple-choice and constructed-response, can provide a more comprehensive and fair assessment of students' mathematical abilities (Reardon, 2018).
3. Reducing language load: Minimizing the language complexity of test items, particularly in constructed-response formats, can help ensure that reading comprehension is not a construct-irrelevant factor influencing performance (Shanmugam & Kanageswari, 2020).
4. Exposing students to higher-order thinking items: Incorporating more HOT items in classroom instruction and assessments can help develop students' critical thinking skills and potentially reduce gender gaps in performance (Shanmugam & Kanageswari, 2020).
5. Training item writers: Providing professional development for item writers on identifying and eliminating gender bias can improve the quality and fairness of mathematics assessments (Alade, Aletan & Sokenu, 2020).

METHOD

In the psychometrics community, item fairness is investigated with differential item functioning (DIF) analysis (Walker, 2011). DIF analysis is an umbrella term for a set of statistical techniques that can be used to compare groups of test-takers by matching on either total test score or estimated ability, and then seeing whether each item measures similarly in all groups (Bandalos, 2018; De Ayala, 2009). The detection of biased items in tests is crucial for ensuring fairness and validity in assessments. Several effective methods have been developed, each with its own strengths and limitations. Here are the most prominent techniques:

1. Analysis of Variance (ANOVA): ANOVA is a statistical method used to compare means across different groups. In the context of detecting biased items, it examines whether the performance on specific test items differs significantly between groups. A significant interaction between item performance and group membership suggests bias. This method is straightforward but relies on the assumption that the items measure a single underlying trait across groups (Wright, Mead & Draba, 1976).
2. Transformed Item Difficulties (TID): The TID approach focuses on the relative difficulty of items for different groups. It assesses whether items are disproportionately difficult for one group compared to another. By transforming item difficulty indices and comparing them across groups, TID can visually represent item bias. This method allows for easy interpretation and can highlight which items require further investigation for bias (Osterlind, 1983).
3. Item Response Theory (IRT): IRT models the probability of a correct response based on both item characteristics and test-taker abilities. By fitting IRT models to data from different groups, researchers can compare item parameter estimates. Significant differences in these parameters across groups indicate bias. IRT provides a robust framework for understanding item functioning and is widely used in educational assessments (Scheuneman, 2005).
4. Chi-Square and Loglinear Models: These models analyze categorical response data to detect patterns that may indicate bias. By examining the responses to distractors in multiple-choice items, researchers can identify items that function differently for various groups. This method can be particularly useful when dealing with complex response patterns (Osterlind, 1983).
5. Qualitative and Quasi-Experimental Methods: Qualitative techniques involve expert reviews and cognitive interviews to understand how different groups interpret items. Quasi-experimental methods compare performance across groups under controlled conditions to assess bias. These approaches provide context and depth to the statistical findings, offering insights into why certain items may be biased (Mellenbergh, 1989).
6. Item Discrimination Index: This is done by finding the discrimination index of the item for both groups. If the discrimination indexes are approximately equal, then the item is probably not biased, but if the values are not approximately equal, such items could be biased (Agi, Hager & Amuche, 2024).

7. Factor Analysis: It can be used to evaluate the internal structure separately for the two groups. If only one factor is found in each group, then the test does not contain bias items, but if more than one factor is found in one of the groups, the test is biased.
8. Rank Order: This is a quick method. Here the test items are ranked in order of difficulty for each of the two groups. If the item rank differs across groups, the test is suspected to be biased.

RESULTS AND DISCUSSION

Research has consistently highlighted the presence of gender bias in mathematics assessments. For example, a study examining the relationship between test item format and gender achievement found that geometry, probability, and algebra items tended to favor male students, while statistical interpretation and multistep problem-solving items generally favored females (Reardon et al., 2018). This suggests that the design of test items plays a critical role in shaping performance outcomes. Further investigation into the sources of gender bias revealed that language comprehension skills could significantly impact performance on mathematics tests, particularly in constructed-response formats. Abedi and Lord (2001) argued that reading comprehension should not be considered a relevant construct in mathematics assessments, as it may introduce bias against students who may excel in mathematical reasoning but struggle with language. McKinley and McCarthy (1984) examined the validity of various item bias detection techniques specifically for mathematics word problems. The research highlighted that different methods yield varying results in identifying biased items, emphasizing the importance of methodological rigor in bias detection. This study demonstrated that empirical comparisons of techniques can inform best practices for ensuring fairness in assessments.

Subkoviak, Mack, Ironson and Craig (2005) administered Mathematics instrument to large samples of blacks and whites. Three popular item bias detection procedures were then applied to the data: (1) the three-parameter item characteristic curve procedure, (2) the chi-square method, and (3) the transformed item difficulty approach. The three-parameter item characteristic curve procedure proved most effective at detecting the intentionally biased test items; and the chi-square method was viewed as the best alternative. The transformed item difficulty approach has certain limitations yet represents a practical alternative if sample size, lack of computer facilities, or the like preclude the use of the other two procedures. Pedrajita (2009) explored gender-related DIF in mathematics assessments and found that computation items, which have a lower language load, still exhibited biases favoring one gender over another. The study suggested that the inclusion of higher-order thinking (HOT) items in assessments could help reduce gender bias by providing a more equitable platform for evaluating student understanding. Moreover, a comprehensive analysis of large-scale standardized tests, such as the PISA assessments, indicated consistent patterns of item format by gender differences across multiple jurisdictions. Male students were found to perform better on multiple-choice items, while female students had higher success rates on constructed-response items (Shear, 2023). This evidence underscores the necessity of considering item format when interpreting gender differences in test scores.

Adedoyin (2010), investigation on gender-biased items in Mathematics examination, he discovered that 5 items were gender-biased out of the 16 test items that fitted the three parameter logistics model (3PL) of IRT statistical analysis. Madu (2012) on analysis of gender related DIF in Mathematics multiple-choice items showed significant gender differential functioning. This implies that the test contained items that measured different things for male and female examinees with the same Mathematics ability. Oribhabor (2019) assessed the unidimensionality and occurrence of Differential Item Functioning (DIF) in the 2017 November/ December WAEC Mathematics test items administered in Edo State, and found that there was occurrence of DIF items in the 2017 WAEC November/ December Mathematics multiple choice test items. Adeosun and Oribhabor (2015) analysed the item parameters of May/June and October/November WAEC 2012 Mathematics Multiple Choice Items, and found that there are presence of differential item functioning in the items. Oribhabor and Omorogiuwa (2014) investigated if items are bias in Edo state BECE 2013 multiple choice mathematic test using differential item functioning approach in relation to gender; and found that 2013 BECE multiple choice mathematics test items functioned differentially across male and female students, in favour of female students.

Kanageswari and Shanmugam (2020) determined the presence of gender Differential Item Functioning (DIF) for mathematics computation items among non-native speakers of English, and thus examining the relationship between gender DIF and characteristics of mathematics computation items. The research design is a comparative study, where the boys form the reference group and the girls form the focal group. The software WINSTEPS, which is based on the Rasch model was used. DIF analyses were conducted by using

the Mantel-Haenszel chi-square method with boys forming the reference group and girls forming the focal group. A total of 988 boys and 1381 girls in form two were selected from 34 schools, with 17 schools located in the Penang island, 12 schools in Penang mainland and five schools in Perak. Some 20 items were selected from the grade eight TIMSS 1999 and TIMSS 2003 released mathematics items. Findings revealed that seven items were flagged as DIF, where two were of moderate DIF and one as large DIF. Two DIF items assessed combined operation from the topics of fraction and negative numbers in the Number domain and the cognitive domain of lower-order thinking skills of Knowing favoured girls. One moderate DIF which assessed higher order thinking skills of applying from the Algebra domain favoured boys. Alade, Aletan and Sokenu (2020) explored the Differential Item Functioning (DIF) of 2018 West Africa Examination Council's Mathematics Objectives Tests Items in Lagos, Nigeria. The research design used for the investigation is a descriptive survey design. The population included all Senior Secondary Three (SS3) students who enrolled for the 2020 West Africa Senior Secondary Certificate Examination (WASSCE) in Lagos State. Multistage sampling procedure was used to select 1334 students from eighteen secondary schools (three schools from each educational district). Three research questions guided the study. The research questions were subjected to item differential functioning analysis using BILOG MG model. Results demonstrated that six items out of the 50 items function differentially in regard to gender. The study uncovered that item analysis using item response theory approach isn't adequate to pass judgment on the nature the test, it is necessary that the item bias is also estimated.

Adewale and Oyeniran (2022) examined the occurrence of item biasness in Lagos State Terminal Unified Mathematics assessment for primary schools pupils in Nigeria. The assessment contained 40 items of multiple choice which was developed by the examination unit of Universal Basic Education Board, Lagos State using primary 5 mathematics curriculum. Primary data of scores of 2018 primary 5 Mathematics 2nd Term Unified Examination was used for this research. A sample of 640 pupils selected through multistage sampling technique was used in the study. Two research questions guided this study. Data was analysed using descriptive and inferential statistics. The results displayed closeness of the mean and standard deviation scores for examinees groups, indicating the examinees have similar ability levels. Out of the items 16 and 17 functions differentially among the examinees based on gender. Adediwura and Asowo (2022) examined the nature of item bias on students' performance in 2017 National Examinations Council (NECO) Mathematics senior school certificate dichotomously scored items in Nigeria. The study adopted an ex-post-facto research design. A sample of 256,039 candidates was randomly selected from the population of 1,034,629 students who took the test. Instrument for data collection was 'Student Results' (SR). Data collected were analysed using the R language environment and an independent t-test. Results showed that the 2017 NECO Mathematics test was essentially unidimensional (-0.28 (<.20), ASSI = -0.31 (< 0.25) and RATIO = -0.31 (< 0.36). Results also showed that the nature of bias statistically encountered was a mean difference in scores bias, indicating that 86% (52 items), 79.1% (34 items), and 96% (56 items) were biased against male students. The study concluded that item bias is a notable factor that affected the validity of the NECO 2017 Mathematics test and conclusions drawn from the scores in Nigeria.

Adebukola (2023) investigated the differential estimate of bias in gender and age in Mathematics anxiety among secondary school students in Oyo State, Nigeria. The sample size consists of 1,500 participants from some selected secondary schools in Ibadan. Ex-post facto design was adopted. The study instrument used was Mathematics anxiety scale. Data were analysed using Mantel-Haenszel procedure for item bias in age and gender. The findings found no significant association between age and mathematics anxiety. A total of 30 items did not exhibit DIF. The study identified that items function differently only with gender. Arıkan (2024) examined bias in one of the well-known mathematics competitions: the Kangaroo Mathematics competition. Determining the fairness of Kangaroo mathematics competition items across gender groups is crucial for creating accurate comparisons and avoiding unintended construct irrelevant bias. To examine the bias, Differential Item Functioning (DIF) analyses were conducted using Logistic Regression, Mantel-Haenszel, and Item Response Theory Likelihood Ratio Test DIF detection methods. After a series of investigations, out of 336 items, it was concluded that these mathematics items were free of DIF and bias across the gender groups.

Ohiri (2024) examined whether the 2020, 2021 and 2022 National Examination Council (NECO) June/July Mathematics multiple-choice questions exhibited uniform and non-uniform gender- related differential item functioning (DIF) in Imo State. A survey research design was employed. The population was made up of all senior secondary school three (SS3) students of 2022/2023 academic session. The number of sampled candidates used in the study was 2,484 students. This comprised 1,178 male and 1,306 female

students. Three research questions were formulated to guide the study. The instruments used for the study were the 2020, 2021 and 2022 June/July multiple-choice mathematics questions set by the National Examination Council (NECO). Each of the instruments consists of 60-items. To detect uniform and non-uniform differentially functioned items by gender, a software called STATA 15 of the logistics regression which is one of the classical test theory methods of DIF detection was applied. The results of the analyses revealed that some items functioned differentially based on gender. Sixteen items (32%) in 2020, sixteen items (32%) in 2021 and twelve items (24%) in 2022, functioned differentially according to the gender of the students.

REFERENCES

Abedi, J., & Lord, C. (2001). The language of assessment: The role of language in assessment. Educational Researcher, 30(7), 4-11.

Adamecz, A., Jerrim, J., Pingault, J. & Shure, N. (2023). Overconfident boys: The gender gap in Mathematics self-assessment. Discussion Paper Series. Retrieved from https://docs.iza.org/dp16180.pdf

Adebukola, K. T. (2023). differential estimate of bias in age and gender on mathematics anxiety among secondary School Adolescents in Ibadan, Oyo State, Nigeria. Journal of Modern European History, 5(6), 557-570.

Adediwura, A. & Asowo, A. (2022). Examining the nature of item bias on students' performance in National Examinations Council (NECO) Mathematics senior school certificate dichotomously scored items in Nigeria. International Journal of Contemporary Education, 5(1), 16-28.

Adedoyin, O. O. (2010). Using item response theory approach to detect gender-biased items in public examinations. Educational Research and Reviews Academic Journals, 5(7), 385-399.

Adeosun, P. K. & Oribhabor, C. B. (2015). Analysis of item parameters of May/June and October/November WAEC 2012 Mathematics multiple-choice items. International Journal of Educational Administration, Planning and Research, 7(2), 305-315.

Adewale, R. & Oyeniran, D. (2022). Examining item biasness in terminal unified Mathematics examination for primary school in Lagos State. The African Journal of Behavioural and Scale Development, 4(1), 10-15.

Agi, C., Hager, A. & Amuche, B. (2024). Evaluation of item bias using differential item functioning (DIF) technique in NECO conducted Economics examination in Taraba State, Nigeria. International Journal of Research and Innovation in Social Science, 8(3), 1776-1790.

Alade, O. A., Aletan, S. & Sokenu, B. S. (2020). Assessing the differential item functioning of 2018 WASSCE Mathematics achievement tests in Lagos State, Nigeria. MB-SDR, 2(2), 8-24.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). Standards for educational and psychological testing. American Educational Research Association.

Aremu, O. A., & Sokan, B. O. (2003). A multicausal evaluation of academic performance of Nigerian learners: Issues and implications for national development. Journal of Applied Psychology, 34(3), 334-342.

Arıkan, S. (2024). Investigating differential item functioning of an international Mathematics competition items across gender groups. Boğaziçi Unıversity Journal of Education, 41-3(1), 53-69.

Bandalos, D. L. (2018). Measurement theory and applications for the social sciences. Guilford Press.

Davidson, M. J., Wortzman, B., Ko, A. J. & Li, M. (2021). Investigating item bias in a CS1 exam with differential item functioning. In proceedings of the 52nd ACM technical symposium on Computer Science Education (SIGCSE'21), March 13–20, 2021, Virtual Event, USA. ACM, New York, NY, USA.

De Ayala, R. J. (2009). The theory and practice of item response theory. Guilford Press.

Federal Republic of Nigeria (2013). National Policy on Education. Lagos: NERDC Press.

Ironson, G., Homan, S. & Signer, B. (1984). The validity of item bias techniques with Mathematics word problems. Applied Psychological Measurement, 8(4). Retrieved from https://journals.sagepub.com/doi/abs/10.1177/014662168400800403

Kane, M. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1–73.

Madu, B. C. (2012). Analysis of gender-related differential item functioning in Mathematics multiple choice items, administered by West African Examination Council (WAEC). Journal of Education and Practice, 3(8), 22-25.

Martinková, P., Drabinová, A., Liaw, Y., 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 (Jun 2017), rm2. Retrieved from https://doi.org/10.1187/cbe.16-10-0307

Mellenbergh, G. J. (1989). Item bias and item response theory. International Journal of Educational Research, 13(2), 127-143.

National Academies (1993). Measuring what counts: A conceptual guide for Mathematics assessment. National Academies Press.

National Council of Teachers of Mathematics (1989). Curriculum and Evaluation Standards for School Mathematics. Author.

National Mathematics Centre, Abuja (2009). Mathematics improvement Programme. Retrieved from www.nmcabuja.org/mathematics_improvement_programme.html.

Niederle, M. & Vesterlund, L. (2010). Explaining the gender gap in math test scores: The role of competition. Journal of Economic Perspectives, 24(2), 129–133.

Ohiri, S. O. (2024). Detection of gender-based uniform and non-uniform DIF of the NECO Mathematics multiple-choice questions in Imo State, Nigeria. International Journal of Innovative Social & Science Education Research, 12(1), 92-103.

Olonode, P. O. (2016). Equating 2014 senior school certificate Mathematics examinations of West African Examinations Council and National Examinations Council in Lagos State, Nigeria. (Unpublished Ph.D. thesis). International center for Educational Evaluation University of Ibadan.

Oribhabor, C. B. & Omorogiuwa, K. O. (2014). Determination of differential item functioning in Edo State Basic Education Certificate Examination (BECE) Mathematics test items (2013) using IRT Analytical Procedure. Nigerian Journal of Health and Kinesiology, 10(1), 222-231.

Oribhabor, C. B. (2019). Assessment of the unidimensionality and differential item functioning in the 2017 West African Examination Council (WAEC) November/ December Mathematics multiple choice test items. Benin Journal of Educational Studies, 25(1 & 2), 1-18.

Osterlind, S. J. (1983). Transformed item difficulties. In Test item bias (pp. 29-38). SAGE Publications, Inc., https://doi.org/10.4135/9781412986090

Pedrajita, A. (2009). Gender-Related Differential Item Functioning of Mathematics Assessments. Journal of Mathematics Education, 2(1), 45-54.

Reardon, S. F., Kalogrides, D., Fahle, E. M., Podolsky, A. & Zárate, R. C. (2018). The relationship between test item format and gender achievement gaps on Math and ELA tests in fourth and eighth grades. Educational Researcher, 47(5), 284–294.

Rover, C. (2005). That's not fair! Fairness, bias and differential item functioning in language testing. Retrieved from http://www2.hawaii.edu/roever/brownbag

Scheuneman, J. (2005). A method of assessing bias in test items. Journal of Educational Measurement, 16(3), 143 – 152.

Shanmugam, S. & Kanageswari, S. (2020). Gender-related differential item functioning of Mathematics computation items among non-native speakers of English. The Mathematics Enthusiast, 17 (1). Retrieved from https://scholarworks.umt.edu/tme/vol17/iss1/6

Shear, B. R. (2023). Gender Bias in Test Item Formats: Evidence from PISA 2009, 2012, and 2015 Math and Reading Tests. Journal of Educational Measurement. Retrieved from https://onlinelibrary.wiley.com/doi/epdf/10.1111/jedm.12372

Subkoviak, M., Mack, J., Ironson, G. & Craig, K. (2005). Empirical comparison of selected item detection procedures with bias manipulation. Journal of Educational Measurement, 21(1), 49-58.

Suurtamm, C. et al. (2016). Assessment in Mathematics Education. In: Assessment in Mathematics Education. ICME-13 Topical Surveys. Springer, Cham. https://doi.org/10.1007/978-3-319-32394-7_1

This Day Newspaper (2024). West African Examination Council. Wednesday, 7th August, 2024. Retrieved from https://www.thisdaylive.com/index.php/2023/08/08/waec-records-80-pass-in-maths-english-language/#google_vignette

Walker, C. M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. Journal of Psychoeducational Assessment, 29(4), 364–376.

Wright, B., Mead, R. & Draba, R. (1976). Detecting and correcting test item bias with a logistic response model. MESA Research Memorandum Number 22. Retrieved from https://www.rasch.org/memo22.htm.