Indonesian Journal of Learning Education and Counseling
Website: https://journal.ilininstitute.com/index.php/IJoLEC
Vol 5, No 1, 2022, pp 01-10
p-ISSN:2622-8068 dan e-ISSN: 2622-8076

Differential Item Functioning and Implications for Testing in Nigeria Education System

Yinusa Akintoye Faremi[1], Kasali Jimoh[2]
[1] Educational Foundations and Management, University of Eswatini, South Africa
Email: akintoyeyinusa@gmail.com
[2] Educational Foundations and Counselling, Obafami Awolowo University, Nigeria
Email: jimoh.bukola@yahoo.com;

Artikel info

Artikel history:
Received: 20-06-2022
Revised: 24-07-2022
Accepted: 21-08-2022
Publish: 25-09-2022

DOI:
doi.org/10.31960/ijolec.V5i1.1689

Abstract. The study analyzes and assesses differential item functioning (DIF) by different demographic groups, particularly gender and cultural groupings, in order to produce appropriate test items. It is essential to examine the extent to which test items work differently among subgroups when selecting test items. This paper is based on modern ways for removing irrelevant parameters and sources of bias of any kind so that a test can produce valid results. As a result, it is recommended that test developers and policymakers evaluate and exercise caution in fair test practice by devoting more effort to more unbiased test development and decision-making. In educational testing, examination bodies should employ the Item Response Theory, and test developers should be aware of test items that may induce bias in response patterns between male and female students or any other sub-group of interest.

Keywords:
Differential Item Functioning;
Test Fairness;
Item Bias;
Item Characteristic Curve;
Item Response Theory.

Coresponden author:
Yinusa Akintoye Faremi
P/Bag 4, Kwaluseni, Swaziland, M201, South Africa
Email: akintoyeyinusa@gmail.com[1]

Introduction

Achievement tests are required by educational institutions in order to determine the desirable features of their examinees. Testing has become one of the most essential criteria by which society judges the quality of its educational system's output. Today's classrooms are made up of a diverse group of students with varying abilities, socioeconomic levels, cultures, and ethnic backgrounds. Teachers' jobs are made more difficult and demanding by the diversity of their students. In a diverse classroom, each teacher is expected to differentiate content to fit the requirements of the students. Students' school performance has consistently been linked to their gender and geographic area. Gender, according to Kanno (2008), is an analytic notion that describes men's and women's sociological roles, cultural duties, and expectations in a specific society or cultural environment. "Gender describes the personality traits, attitudes, behavior, values, relative power, influence, positions, and expectations (femininity and masculinity) that society ascribes to the two sexes on a differential basis. Ezeh (2013). As a result, gender is a psychological concept as well as a cultural construct used by society to distinguish between male and female roles, behavior, mental, and emotional characteristics. One of the reasons for examining test takers in schools is to provide test results which are often used in making important decisions such as selection, promotion and certification. Based on test scores, schools choose who are to be promoted, external examining bodies decide who are to be certificated, higher institutions decide who are to be admitted and for which course, and recruiting organizations choose who are to be selected. Since the decisions made on test scores are extremely significant to the individual and the public, these should reflect the most accurate estimates of their abilities and skills.

Over time, tests have been used to evaluate students' academic performance. The aim of tests in the academic system is to determine the traits and attributes of test takers. Terminal examinations are commonly conducted in Nigeria from one stage to the next, especially at the primary and secondary school levels. These examinations are being conducted by different examination bodies like West African Senior Secondary Certificate Examination Council (WASSCE), National Examinations Council (NECO, State Ministry of Education), National Business and Technical Examination Board (NABTEB), and so on. Different test takers of different levels of ability with different languages, cultures, sexes and religions are made to undertake these examinations. Considering the significance of examinee ratings in making important decisions, educational evaluation aimed at enhancing the decency of tests or assessments across subgroups of test takers is critical. The examination contains a collection of structured tasks to which candidates or test takers are allowed to respond individually; the results of this examination evaluate each candidate to provide a statistical correlation to their abilities (Nworgu, 2011).

Concept of Differential Item Functioning

Differential item functioning had previously been referred to as "item bias" in the literature since it causes one group to have a higher scale score than the other Differential item functioning occurs when there is existence of some irrelevant elements present in an item which causes differential performance for individuals of the same ability but from different ethnic, gender, type of school attended. (Ogbebor & Onuka, 2013). When examinees with the same ability have different probability of responding well to an item based on group membership, this is referred to as differential item functioning (e.g., male or female). Based on the movement of disparities between groups over the ability spectrum, there are two types of DIFs (i.e., total test score). When one group consistently performs better or worse than the other across the ability spectrum, uniform DIF occurs. The heading of the differentiation moves throughout the ability continuum, assuming that the group member and the ability are linked in any situation. This is referred to as non-uniform differential item functioning in this scenario. A substantial amount of unequal item functioning in test bits demonstrates a lack of construct validity in test items. Additional nuisance constructions that function differently from one group to the next are measured by the items with differential item functioning. Disturbance constructions can jeopardize the most accurate assessment of a subgroup's output (Park, Pearson & Reckase, 2005). Differential item functioning, according to Doolittle and Cleary (1987), is a condition in which the likelihood of correctly responding to a question is linked to group membership among examinees of similar skill levels. In terms of probability, Warm (1978) used the following equation to characterize differential item functioning:

$Pa\ (\theta = K) \neq Pb\ (\theta = K)$……………….equ (1)

In the equation above, A and B address two subgroups, whereas addresses ability or latent characteristic, which would be equal to K from every group membership. The requirement states that an item should be differentially working if the chance of people from bundle A ability (K) getting an item right is not equal to the probability of people from group B ability (K). When various subgroups that are composed in terms of the fundamental estimation alter their excess on an assistant estimation, such as details on the material to the degree that the elements are definite, DIF arises. DIF evaluations are aimed at detecting objects that are influenced by assistant estimations or, more shockingly, that behavior phenomenal, additional points in various subgroups. This term is now only used when objects have been interpreted as differentially operating by quantifiable strategies, and the protection can be blamed for creating unimportant properties of the object (Lam, 1995). Furthermore, Differential item Functioning depicts differences in individual functioning across items and groups of people. Differential Group Functioning (DGF) is a form of DIF that shows the differences in functioning between item classes and individual classes, as well as differential item functioning.

Differential item functioning occurs when people from various groups (usually gender identity or character) with the same latent qualities (ability/latent) have differing probabilities of responding to a questionnaire or survey (DIF). Differential Item Functioning is demonstrated when and only when people from various social groups with a similar secret verifiable ability have a varied likelihood of answering a question. Factor relates to test items that establish demands that are different from those given by the test designs, such as changing meanings or recommendations for people from different groups or semi-groups. Differential item functioning occurs when the difficulty level (b), discrimination level (a), or lower asymptotes (c) of a test item vary between groups of test takers, as defined by item response theory (IRT). If a subset of the general being tested responds differently to a few test items, it suggests the items are frequently more difficult for one group than the other. 'Differential item functioning happens when an item is not equally difficult or popular in maximal performance tests for groups that have been matched in terms of the construct being measured' (Lincare, 2011).

The society (uban/rural) where an individual find himself or herself has a significant impact on academic performance over the course of one's normal daily life. People who grow up in a wealthy community gain higher levels of insight than those who grow up in a poor society. Larger cities are exceptional, with learning centres, skilled teachers, excellent paths, and exceptional communication networks, allowing them to stand apart from their national counterparts where such resources are insufficient or by any means inadequate.

According to Akubuiro (2002), referenced by Anagbogu (2009), metropolitan learning environments provide more access to socio-cultural and economic facilities and services, resulting in high-performing learners. Rural learners who have not yet been exposed to these beneficial experiences and crucial physiological functions find it difficult to overcome any hurdles along these lines, resulting in surprising results in their altered subjects. Understanding the relative strength and weakness of the examinee groups on the various skills and talents that the test items measure requires identifying the causes of DIF. Item content, item type or format, item context, content, and cognitive factors linked with objects are all plausible origins for such trends. By examining the statistical evidence of item level DIF in light of such item qualities, it may be feasible to obtain significant insight into the likely causes of DIF. In practice, items that exhibit significant DIF are not always removed from future tests, but they are among those that must be thoroughly examined before being used again. When people with the same skill level but belonging to different groups are seen to have differing chances of answering correctly to an item, this is known as differential item functioning That is, after they are matched on the ability that the test was designed to measure, systematic disparities in performance between different groups of examinees are detected (French & Finch, 2010).

Techniques for Detecting Differential Item Functioning

In this section, we review the most commonly used statistical methods that have been developed to detect DIF (Magis et al., 2010). We focus on methods for tests with dichotomous items, which include binary items graded as true (1) or false (0), or as correct (1) or incorrect (0), on multiple-choice or free-recall tests. Methods for detecting DIF on other types of items (e.g., those graded on a rating, ranking, or partial-credit scale) are similar but beyond the scope of this paper. Generally speaking, statistically detecting items exhibiting DIF requires that we match students on relevant knowledge (e.g., using their total scores on the assessment being evaluated as an estimate of ability, or latent trait), and then test whether students who are matched for ability but from different groups perform similarly on a given item.

The methods for detecting DIF vary depending on how students are matched. Classical methods (e.g., Mantel-Haenszel statistic and logistic regression) match students based on their total scores; methods based on item response theory (IRT) models, such as the Wald χ2test (also known as Lord's test) and Raju's area test, consider student ability as a latent variable estimated together with item parameters in the model

Generally, IRT methods are computationally more demanding and require larger sample sizes. However, IRT methods are more precise than others, because they more accurately estimate the latent trait instead of using total score as the proxy.

In this paper, we discuss methods based on IRT which more accurately estimate both item characteristics and student abilities but require relatively larger sample sizes.

Item response theory as methods for detecting differential item functioning

Item response theory (IRT) is an effective approach to examine differential item functioning. Item bias or differential item functioning, in which one group responds differently to an item than another group, is an essential tool in item analysis. Researchers can use information about the item's location or the underlying trait to identify or determine the levels not assessed or explicitly measured by the instrument. There are various graphical illustrations to examine the item's location along the underlying latent trait (θ) after estimating item properties. DIF occurs when an item performs differently for respondents in different groups. In other words, members from different populations who have equivalent levels of a latent feature (e.g., physical functioning) have a varying chance of answering to an item. Differential item functioning items pose severe danger to the validity of instruments used to assess members of various populations or groups. Instruments with such items may have lower validity for between-group comparisons since their score could indicate a variety of characteristics other than those measured by the scale. The possibility that some test items are unfair to one subgroup or another has become a source of worry for both test developers and test users. The most extreme definition of item and test bias states that a test is biased if the means of two groups of interest diverge. The obvious flaw in this definition is that other variables, in addition to item bias, have a role in these disparities.

Item response theory is the most common framework in which differential item functioning is characterized (IRT). The item trace line allows you to compare the responses of two different groups to the same item, such as reference (e.g., control) and focal (e.g., treatment). Item parameters are expected to be invariant to group membership in the context of IRT (in contract to classical test theory where parameter estimates and statistics vary with the sample being measured). As a result, the difference between the trace lines, calculated independently for each group, demonstrates that at the same level of the underlying trait, respondents from the reference and focal groups have differing probability of approving the item. DIF is defined as when the conditional probability, p(x), of an item's correct response for the same level on the latent variable differs between two groups (Camilli & Shepard, 1994).

Dichotomous data, generally examined for one or both of two types of DIF. Uniform differential item functioning (DIF) tends to advantage one group over the other across the whole setoff ability. Non-uniform differential item functioning (DIF) however, exists when there is an interaction among membership within a group and ability level (Narayanan & Swaminathan, 1996). Within item response theory (IRT), dichotomous item responses are often modelled using some variant of the general curvilinear three parameter logistic (3-PL) model shown.

$$P(\theta)=(\ c+1)\,\frac{e^{ai(\theta-bi)}}{1+e^{ai(\theta-bi)}}\ \ldots\ldots\ldots\ldots\ldots\ldots\text{equ (2)}$$

In this model, ***ai*** is the parameter for the item discrimination, ***bi*** is the parameter for the item's difficulty, and ***ci*** is the item's pseudo-guessing parameter where ***θs*** is the parameter for examinee's capacity. Here, the probability of an examinee's correct response to the item is dependent solely on the three item parameters and the one-individual parameter and, disregarding error, should be the same regardless of how the examinees are grouped (Embretson & Reise, 2000). Inside this structure, uniform differential item functioning is seen when two or more groups vary on item difficulty parameter after connecting on ability. Be that as it may, if the group differ, either all items being equal or also, on the item discrimination parameter, at that point non-uniform DIF is obvious (de Ayala, 2009).

Differential Item Functioning is of two types of uniform differential item functioning and non-uniform differential item functioning. More particularly, uniform differential item functioning happens "when a group performs better than another group on all ability levels" (Karami, 2012), group membership does not interact with level of ability. However, non-uniform differential item functioning happens in situations that "members of one group are favored up to a level on the ability scale and from that point on the relationship is reversed", and an interaction exists between group membership and level of ability (Karami, 2012). There are at least two groups when running differential item functioning analyses, classified as either focal or reference groups. The focal group relates to the minority group while the reference group pertains to the majority group (Cuevas & Cervantes, 2012). Regardless of the method used to detect differential item functioning, the focal group's item responses are compared to those of the reference group in order to identify items bringing about different performance of the two groups.

A graphical portrayal of uniform and non-uniform DIF can be acquired utilizing separate item characteristic functions for every one of the two gatherings. Figure 1 shows an illustration of uniform DIF. In this figure, the two item characteristic functions (ICF) values contrast just in difficulty parameter. The difficulty parameter in this example has a value of -0.5 for Group 2 (the reference group) and a value of 0.5 for Group 1 (the focal group), indicating that this item is harder (or takes more of the trait to answer correctly) for Group 1 than Group 2. Since this is the only parameter that varies between the groups for this item, Group 2 has a higher probability of correctly answering the item than Group 1 across every level of ability.

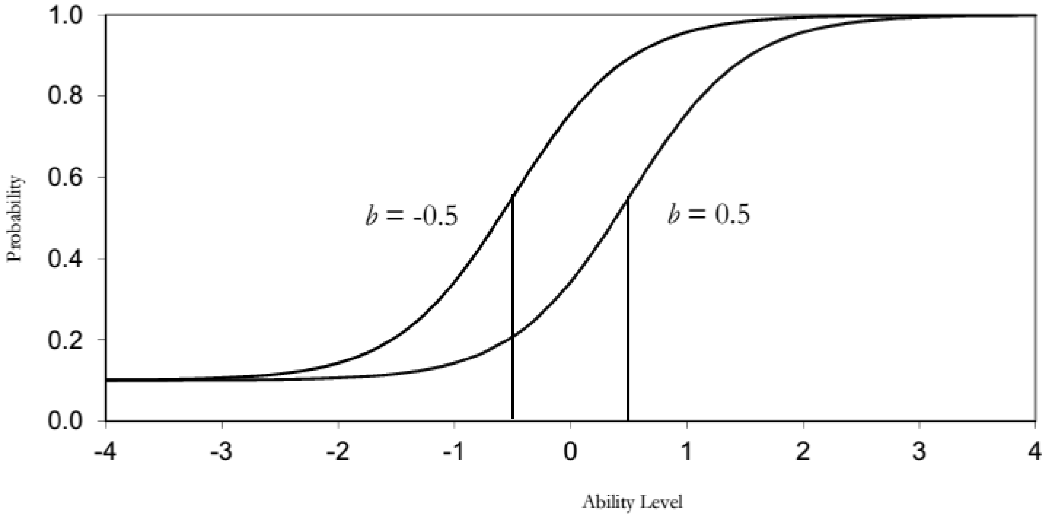

Figure 1. Probability curves for two groups on an item that displays uniform DIF.

The item characteristic curves (ICC) displayed will not meet except that they asymptote to the same values (in this example at .10 and 1).

Figure 2 is an example of non-uniform DIF where not only is there a difference between groups in probability in correctly answering the item but the group having the advantage changes at some point within the ability range. Here, one group has a higher probability of correctly answering the question at end of the scale and the other group has a t the other (Walker, 2001). In this example, the item characteristic function values differ in terms of guessing parameters, as well as the difficulty parameter. The item characteristic curve in this example show that, at the lower levels of ability, examinees in Group 1 have higher probability of correctly answering the item than those in Group 2, where this trend is reversed at the higher end of the ability scale. Some researchers have termed non-uniform.

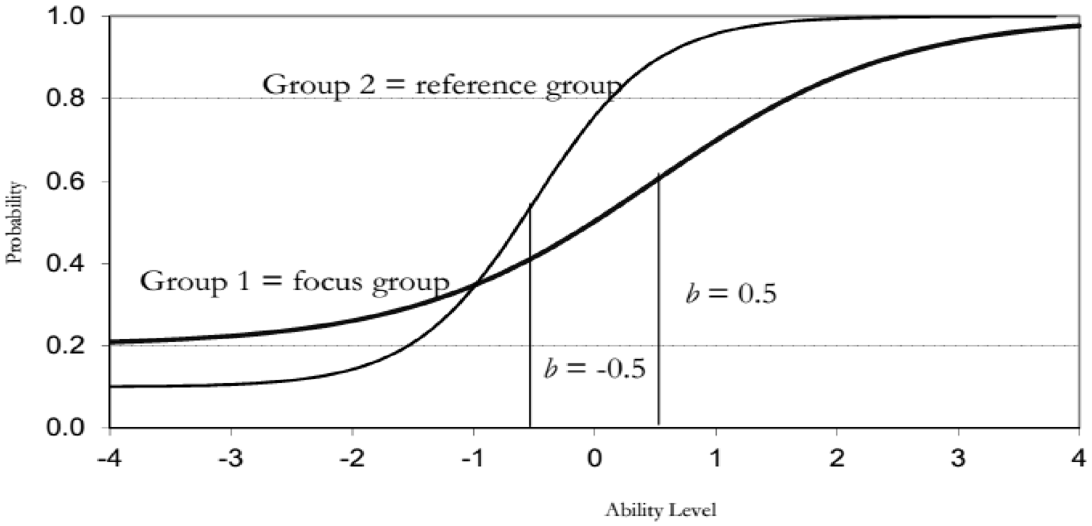

Figure 2. Probability curves for two groups on an item that displays non-uniform DIF.

When using the item characteristic curve to evaluate our test item under item response theory, different types of curves can be derived for each item and they can easily be interpreted foe more objective item analysis in our validation processes. It is also necessary to note that

- The steeper the curve, the better the item can discriminate
- The flatter the curve, the less the item is able to discriminate since the probability of correct response at low ability level is nearly the same as it is high ability levels
- The steepness of the curve in its middle section indicates the rapidity with which the probability that examinee responding to the question correctly changes as a function of ability

The location of the curve along the horizontal axis (as defined by the point at which the 0.5 probability level bisects the horizontal scale) indicates the difficulty of the item.

Figure 1-7 shows the item characteristics of the items of the 2016 NECO Mathematics test. Jimoh (2021)

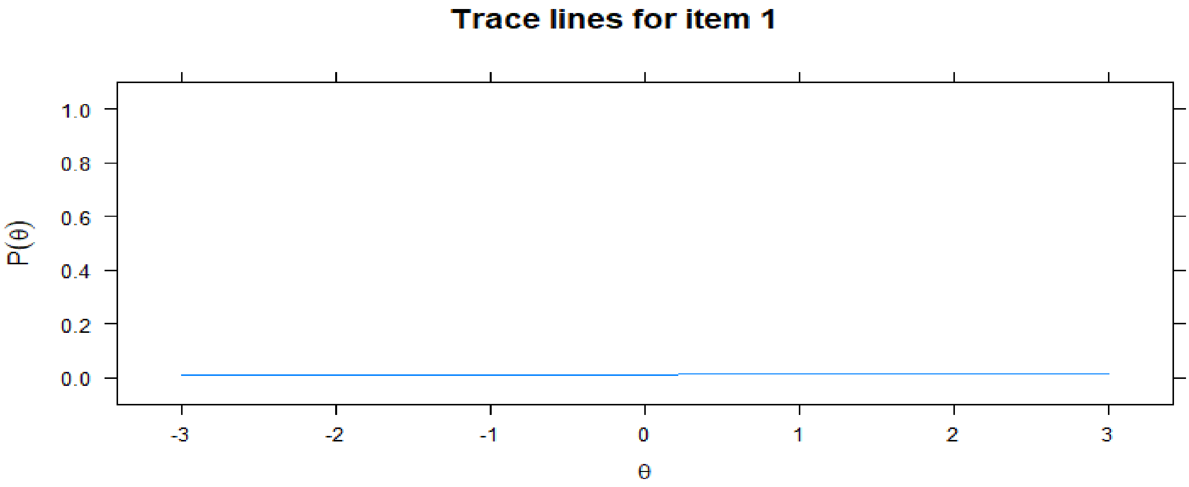

Figure 3. Item Characteristics Curve for Item 1

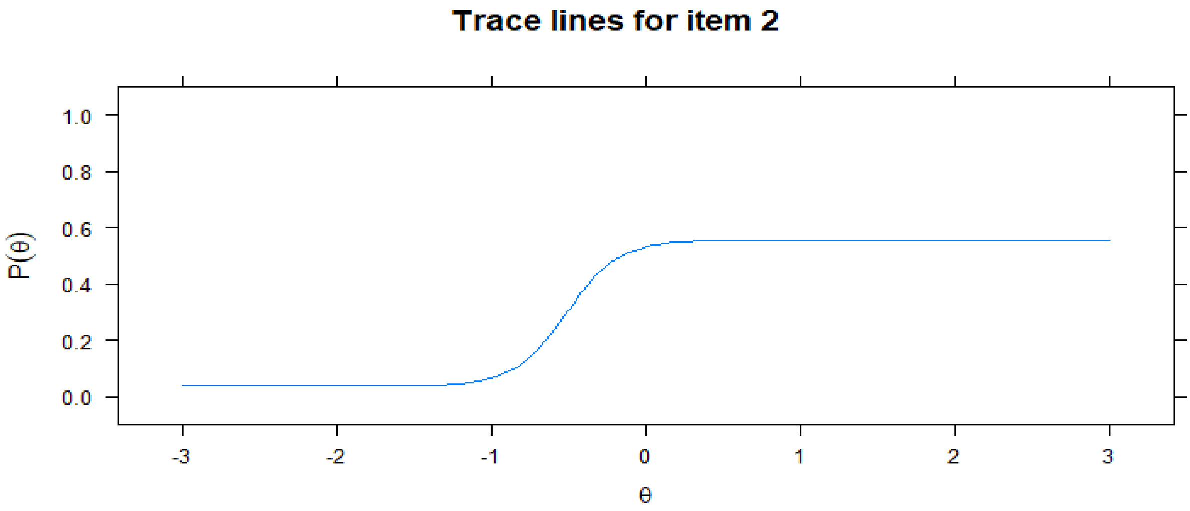

Figure 4. Item Characteristics Curve for Item 2

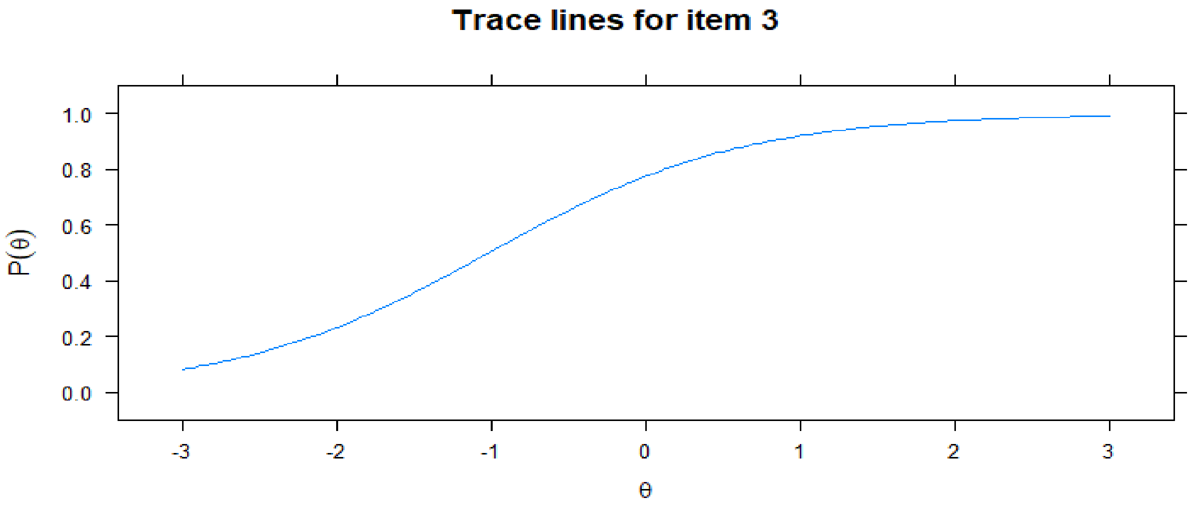

Figure 5. Item Characteristics Curve for Item 3

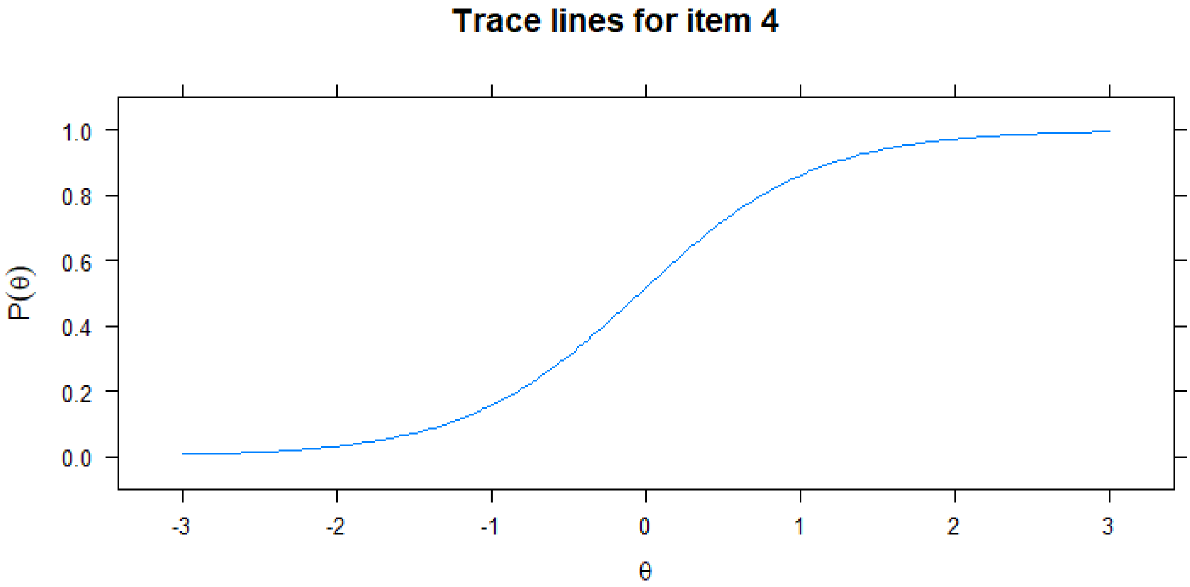

Figure 6. Item Characteristics Curve for Item 4

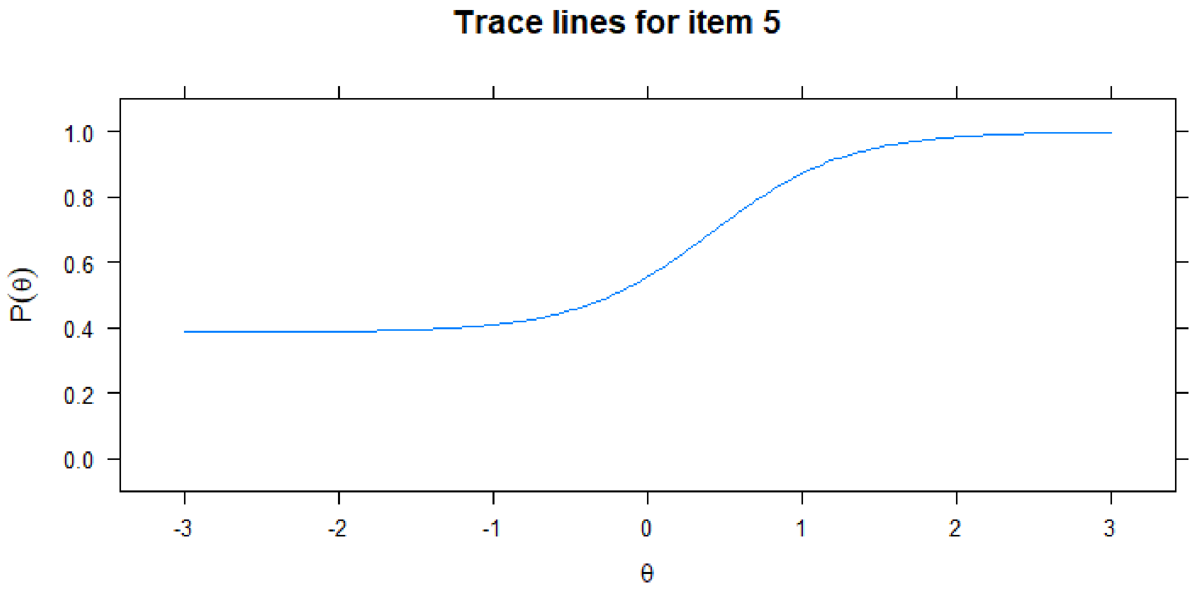

Figure 7. Item characteristics curve for item 5

Differential item functioning can detect item bias if and only if the following assumptions are true:

a. That the test items are all measuring the same thing. What if there are numerous characteristics at play? A math examination, for example, may include items in both numeric and essay formats. A tester with a poor math talent but a high reading skill may be able to properly answer tough math problems expressed in numeric form but not easy questions expressed in essay form. Because practically all assessments require reading skills, this dual-trait condition is inescapable. DIF, on the other hand, may not be beneficial when an ability incorporates numerous features other than content and language, such as A and B. On A, the test item may be biased against men, but on B, it may be biased against women.

b. That the entire test is fair, with only a few items being biased. As a result, the overall score is evaluated to divide testers into high and low ability groups. What if, on the other hand, more than half of the items are biased?

c. That the abilities are spread equally across groups. As a result, rather than underlying group discrepancies, some test score variations are due to unfair questions. Those who question this assumption are sometimes labeled as "racists," "sexists," or "narrow-minded." However, rather than being settled by cultural relativism, whether particular groups have intrinsic advantages on certain tasks should be subjected to scientific inquiry.

Implication of differential item functioning or biased item in Educational System

Bias can result in systematic errors that distort the inferences made in any selection and classification. As mentioned earlier, there exist a number of examination bodies in Nigeria and these bodies cater for candidates of various backgrounds all over the country. Candidates who participate in the examinations conducted by these examination bodies are in different settings and therefore differently toned for personal and environmental reasons. As a result of this, the problem of test item bias cannot be ruled out in these examinations. It is expedient that the examining bodies examine the degree of bias in their examinations. It has been claimed that some of the national examinations unfairly favour examinees of some particular groups eg, cultural or linguistic groups to the extent that it is now believed that a particular section of the country perform most woefully in these national examinations.

A critical look at the perception of people on such national examination in Nigeria indicates the serious nature of differential item functioning. For a test to be free from bias, it must be unidimensional. Unidimensionality is the assumption that an item is intended to measure a single attribute or skill for all examinees. The assumption of unidimensionality is the most complex and most restrictive assumption of item response theory. In general, unidimensionality means that the items measure one and only one area of knowledge or ability. Lumsden (2003) provides an excellent method for constructing unidimensional tests. "Item response theory provides a test of item equivalence across groups. We can test whether an item is behaving differently for blacks and whites or for males and females", for example. Jimoh (2021) carried out a study on gender and culture-related differential item functioning in 2016 National Examinations council Mathematics multiple choice questions in Nigeria. He discovered that the items of 2016 NECO Mathematics multiple choice test functioned differentially among Hausa/Fulani, Igbo and Yoruba cultural environments.

Conclusion and Suggestion

Test results are utilized to determine or take judgments and recommendations in our schools, biased items and differential item functioning are particularly relevant in our educational systems. Differential item function is concerning because, once examinees have been matched based on their interest in a psychological characteristic, test-takers from different demographic groups, such as gender, have varying chances of passing a test item. When learning opportunities are not evenly distributed, test takers from different categories may be expected to differ in ability at times. In these circumstances, item impact rather than item bias is frequently employed to explain the outcome. DIF analysis is one of the most effective methods for removing extraneous elements and sources of bias from a test so that valid and reliable results can be obtained.

Test experts and developers must always put in place quality control mechanisms so that they can deliver high-quality test item; educational measurement experts in Nigeria should rise to the challenges placed by the measurement community and be fully aware of the usefulness of IRT in constructing and scoring of tests or examinations and examination bodies should organize training for item developers on the construction of valid, reliable and fair test especially in the area of DIF. In addition, items flagging DIF should be revised, modified or eliminated from the test

REFERENCES

Akubuiro, I. M. (2002). *Self-concept, attitude and performance of senior secondary school students in physical science subject in Southern River State, Nigeria* (Unpublished master's thesis). University of Calabar, Calabar, Nigeria.

Anagbogu, G. E. (2009). *Analysis of psychometric properties of WAEC and NECO Examination instruments and students' ability parameters in Cross River State-Nigeria.* An unpublished PhD thesis Department of Educational Foundations, University of Calabar.

Baharloo, A. (2013). Test fairness in traditional and dynamic assessment. *Theory and Practice in Language Studies, 3*(10) 1930-1938. doi:10.4304/tpls.3.10.1930-1938.

Camilli, G. (2006). Test fairness. In R. Brennan (Ed.), Educational measurement, 4th ed. (pp. 221–256), Westport, CT: American Council on Education and Praeger.

Cuevas, M., & Cervantes, V. H. (2012). Differential item functioning detection with logistic regression. *Mathematics and Social Sciences, 50*(199), 45-59.

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

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, N J

Ezeh, A. (2013). Is gender a factor in mathematics performance among Nigeria pre-service teachers? *Sex Role, 51*(11&12), 749 -754.

French, B. F., & Finch, W. H. (2010). Hierarchical logistic regression: accounting for multilevel data in DIF detection. *Journal of Educational Measurement, 47*(3), 299-317.

Jimoh, K. (2021). *Gender and culture-related differential item functioning in 2016 National Examinations Council Mathematics multiple choice questions in Nigeria*. (Unpublished doctoral thesis).

Faculty of Education, Obafemi Awolowo University, Ile-Ife.

Kanno, E. E. (2008, August 13). The deprived Cross Riverians. *The Nigerian Chronicle.*

Karami, H. (2012). An introduction to differential item functioning. *The International Journal of Educational and Psychological Assessment, 11*(2), 59-76.

Lam, T. C. M. (1995). Fairness in performance assessment. *ERIC digest* (online). Available: http://ericae.net/db/edo/ED391982.htm (ERIC Document Reproduction service No. ED 391 982)

Li, H. H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika, 61,* 647-677. doi:10.1007/BF02294041

Linacre, J. M. (2011). *The effect of misfit on measurement* [Paper presentation]. Eighth International Objective Measurement Workshop, Berkeley, CA.

Ogbebor, U & Onuka, A. (2013). Differential item functioning method as an item bias indicator.*International Research Journal*. 4(4)367 – 373.

Park, H. S., Pearson, P. D., & Reckase, M. D. (2005). Assessing the effect of cohort, gender, and race on differential item functioning (DIF) in an adaptive test designed for multi-age groups. *Reading Psychology, 26*(1), 81-101.