INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN PROGRESSIVE EDUCATION & DEVELOPMENT

OPEN ACCESS JOURNAL

Vol 1, Issue 1, (2012) E-ISSN: 2226-6348

Issues in Test Item Bias in Public Examinations in Nigeria and Implications for Testing

Dr. Emaikwu, Sunday Oche

Senior Lecturer, College of Agricultural & Science Education, Federal University of Agriculture, Makurdi Benue State Nigeria

Email: emaikwuoche@yahoo.com

To Link this Article: http://dx.doi.org/10.6007/IJARPED/v1-i1/10942 DOI:10.6007/IJARPED/v1-i1/10942

Published Online: 24 March 2012

Abstract

Various criticisms have been leveled against psychological testing. A large proportion of the criticism pivots on fairness of test to the various identifiable groups taking the same test. This article discusses the meaning, types and sources of test item bias as well as different methods of detecting it in a testing situation. One of the implications of test item bias in Nigerian educational system is that it blurs the essence of testing which is to reveal the latent ability of examinees. Test item bias also affects the vital psychometric properties of measurement results in terms of validity and reliability. It is recommended that examination bodies should construct test items in such a manner that items are free from writing errors such as wordiness, irrelevancy, offensiveness, and excessive stimulations, so that when an inadequacy exists between groups' examination scores, the disparity will be attributed to true differences in whatever the test purports to measure in the examinees. The paper emphasizes that educators should take more cognizance of the possibility of test item bias in a testing situation and with this kind of effort, candidates from educationally disadvantaged areas and low socio-economic status would be certain to be fairly treated.

Keywords: Bias, Public Examinations, Implications.

Introduction

Educational institutions are expected to conduct achievement tests to be able to establish the desired characteristics of their examinees. Testing has become one of the most important parameters by which a society adjudges the product of her educational system. Testing has always been an important part of the school system that even the habitual absentees normally turn up to school and present themselves for testing on examination days. The essence of testing is to reveal the latent ability of examinee. Testing has been fully accepted in most modern societies as the most objective method of decision making in schools, industries and government establishments. It is now used for admission, recruitment, promotion, placement, evaluation, guidance, research and teaching purpose among others (Emaikwu, 2011). Though accepted in most societies as the most objective method of decision-making, nonetheless, the use of test has sparked off some concerns among the members of the public in recent years. These concerns have tended to erode people's faith in the power and efficacy of tests. The most serious allegation voiced so far against testing

pivots around the social issues that test may show culture or class bias (Anastasia & Urbina, 2006). Test items designed to test abilities and interests of the children in the high and middle class may not make much sense to children in the low socio-economic class. The issue of bias in testing is currently appearing in public forums including courts of law, and decisions are being made that have an impact on critical issues such as who shall be educated and who shall be employed (Berk, 2007).

The interpretations that are put on the results of tests create a lot of problem. Sometimes, many candidates are disqualified as a result of problem of making classic distinction between the aptitude and achievement tests. If a test is an achievement test and is interpreted as such, then a high score means that a course of instruction has been successfully conveyed and that further educational effort in that area is unnecessary. On the other hand, a low score means that additional educational effort perhaps more of the same or of some other kind is called for, since the achievement has not occurred. But if the test is seen as a test of pure aptitude test, in that case, instead of measuring accomplishment, the intent is to measure the capacity for accomplishment (Nunnally, 2008). Thus a high score portends very well for the test taker but a low test score may be interpreted as an indication that there is insufficient capacity on that test taker's part to achieve; therefore any additional educational endeavor would be an effort in futility or a wasteful one. In this distinction between the interpretation of achievement and aptitude tests lies a social decision of great significance. If a test performance is low and it is seen as index of achievement, then there is a pressure to increase the application of the society's educational resources to improve that achievement. If the test result is seen as an index of aptitude then the same low test score may be seen as a justification of withdrawal of educational resources. When this misinterpretation is brought to bear in the school system then there is the possibility that some examinees will be unduly treated and hence whenever this situation occurs, then test bias is presumed to exist in the measurement of ability. Test bias in measurement has become a heated, complex and pronounced issue in the western countries and most developing countries are also becoming conscious of the concept even though there is low use of psychological test in those developing nations (Joshua, 2005).

In Nigeria, there exist a number of national examination bodies and they include National Examination Council (NECO), West African Examination Council (WAEC), National Business and Technical Examination Board (NABTEB), and Joint Admission Matriculation Board (JAMB). 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 has been claimed that some of the national examinations unfairly favor examinees of some particular groups e.g., 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 item bias. A test item that is not unidimensional is of course, not free from bias. For example two items designed to assess multiplication skill in mathematics could be as follows: (i). what is 6×7? (ii). what is the product of six and seven? Item (i) requires only knowledge in mathematical operations, while item (ii) requires for its solution, a specific

amount of reading competence as well as knowledge of mathematical operations. When different attributes are being measured as in item (ii), the issue of item bias enters into consideration if such item is administered to two different groups and the responses of one of the groups are dependent on the secondary skill. This type of item measures different types of skills among different groups. If the test makes the members of one group look worse than their attainment would actually be on the job or in the classroom, the test is said to be biased against that group. The same notion of bias is applied to school achievement test; if for instance, children in one group consistently receive lower scores than would be expected from their observed classroom performance. In Nigeria, the national examinations are the likely examinations that may suffer from bias problems.

Examination bodies often carry out empirical verification for detecting biased items in their respective examinations in order to redeem and exclude items found to be biased so that all the examinees can be assured of equity in the examination and also to ensure that the ability of examinees are reliably assessed Examination bodies are expected to construct test items in such a manner that test items are free from writing errors such as wordiness, irrelevancy, offensiveness, and excessive stimulations, so that when an inadequacy exists between groups' examination item scores, the disparity will be attributed to true differences in whatever the test purports to measure in the examinees (Dibu-Ojerinde, 2006). As educators take cognizance of the possibility of test item bias in national testing situation, candidates from educationally disadvantaged areas and low socio-economic status would be certain to be fairly treated.

National examination bodies often over-predict or under-predict some candidates from certain states during the selection exercise, to the extent that some examining bodies have different policies of awarding the final grade to examinees. For instance, JAMB has accepted different cut-off points for selecting candidates into Nigerian tertiary institutions based on merit, catchment area, educationally disadvantaged states and institutional discretion. Nigerian as a nation is a heterogeneity setting and there is an assumption that human development is a process dependent upon interaction between inherited qualities and environmental forces. Often times, the language of the test item could be so apt to be offensive to members of a particular subgroup. Sometimes, test items are constructed in a manner that some elements of it could offend examinees on ethnic, sexual, cultural, religious or socio-economic grounds. For instance, a test item on Social Studies that described "Alhaji Adeleke Adebayo as the leader and mastermind of a bloody riot during the house of assembly election in Makurdi" is a biased item. It is obvious that some Muslim candidates would be offended because of the implication that an '*Alhaji*' led the bloody riot in the first instance. On this term too, Yoruba candidates may feel offended that their kinsman is involved in the heinous riot. The candidates from Benue State and those who have relation in Makurdi town may concentrate on the issue in the item instead of the examination.

The procedures employed in the administration of the national examination could be sources of bias. The actual administration of the examinations constitutes a complex interaction among examiners' variables, examinees' variables and situational variables. In Nigeria, the concept of test item bias is among the topical issues of concern and has become a daily issue of national discourse even at the legislative assembly. The Nigerian senate in 2010 summoned the then minister of education to the senate chamber to explain why the massive failure

occurred in that year's national examination; the issue of test item bias and test-wiseness featured prominently among other reasons given for massive failure in some sections of the country. During post unified tertiary matriculation examination exercise, many candidates often complain of biasedness in the testing process as some tertiary institutions are accused of setting '*local and irrelevant*' questions extraneous to candidates' areas of specialization. Moreover, it is obvious that post unified tertiary matriculation examination conducted by many tertiary institutions is not based on any known syllabus and hence this could serve as a major source of bias in the selection process.

Psychological testing as a procedure and psychological tests as instruments have come under various criticisms since testing began in the school system. A large proportion of the criticisms revolve on fairness of test to the various identifiable groups taking the same test. Measurement experts have begun to intensify research work in this interesting, sensitive and sentiment-laden issue especially now that universities and other tertiary institutions conduct aptitude tests for students to complement JAMB admission procedures. Sequel to these submissions, the article discusses issues in test item bias in public examinations in Nigeria. It specifically examines the meaning, types and sources of test item bias as well as different methods of detecting it in a testing situation. Also the implications of test item bias in Nigerian educational system especially at post primary school level were highlighted.

The Concept, Sources, and Types of Test Bias in the Measurement of Ability

The issue of fairness is what critics labeled "bias" in testing. When the whole test is the unit of concern, then *"test bias"* is the issue to be examined whereas when an individual item is the unit of concern, then "*item bias"* is the concept of focus. A more decorated term for item bias has been formulated namely "differential item functioning". Differential item functioning (DIF) occurs when examinees from different groups show differing probabilities of success on the item after matching on the underlying ability that the item is intended to measure (Zumbo, 2009). Item bias occurs when examinees of one group are less likely to answer an item correctly than examinees of another group because of some characteristic of the test item or testing situation that is not relevant to the test purpose. An example of hypothetical item bias is

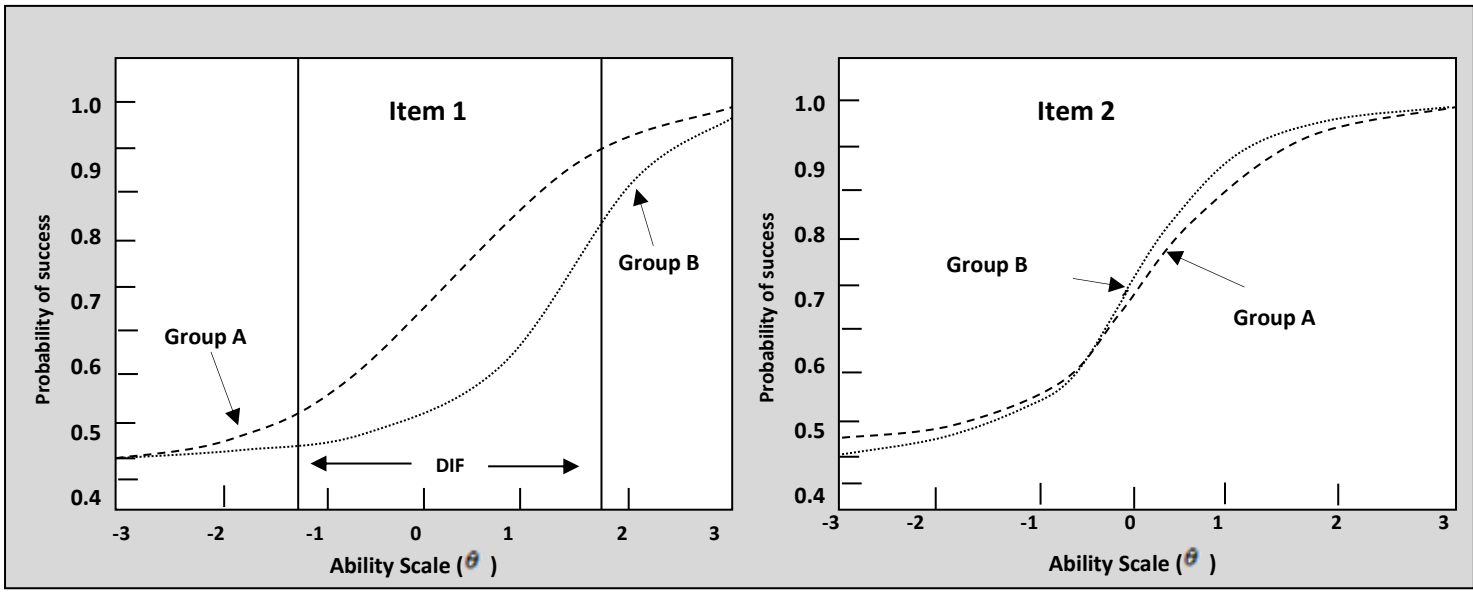

Fig. 1: Item Characteristic Curves (ICC) for two items, illustrating large and small amounts of differential items functioning (DIF). (Graph adapted from Anastasia & Urbina, 2006)

If the ICCs are identical for each group, or very close to identical, it can be said that the item does not display DIF as seen in item 2, but if the ICCs are significantly different from one another across groups as observed in item 1, then the item is said to show DIF. By comparison, it can be seen that for item 1, the ICCs for groups A and B are quite dissimilar, while for item 2, they are closely similar. Item 1 therefore gives an example of an item that displays substantial DIF with a very large area between the two ICCs. That a test item is not biased is an important consideration in the selection and use of any psychological test, that is, it is essential that a test is fair to all applicants, and is not biased against a segment of population taking the test items. In many cases, test items are biased due to the fact that they contain sources of difficulty that are irrelevant or extraneous to the construct being measured, and these irrelevant factors affect performance. Perhaps the item is tapping a secondary factor or factors over-and above the one of interest. When items have the same construct validity for all examinees in a population, examinees of comparable ability may have the same chance of getting the item correct (Berk, 2007).

The kinds of bias that may be encountered in tests ranges widely and they include sex bias, religious bias, geographic bias, linguistic bias and racial ethnic heritage bias. If the test makes the members of one group look worse than their attainment would actually be on the job or in the classroom, the test is said to be biased against that group. The same notion of bias is applied to school achievement test; if for instance, children in one group consistently receive lower scores than would be expected from their observed classroom performance Four types of test bias could easily be encountered in the process of testing and they include content bias, atmosphere bias, and bias in use-social consequences.

Content bias occurs when the content of the test items gives a systematic advantage to a particular group of test takers. Usually the bias reflects differences in the opportunities to learn the material tested. Test items may be biased and unfair to the members of any group if they have not had the opportunity to learn the material. However, if members of various groups have had equal opportunities to learn the test contents, any observed differential performance may not be persuasive evidence of content bias. Moreover, atmosphere bias could arise as a result of the testing conditions on the examinees' performances. It could emanate from the type of motivation elicited, factors related to the examinees-testers interaction, and factors in the evaluation and scoring of responses. The goal in testing is to minimize any possible test condition effects and this is usually accomplished by using standard testing conditions.

According to Camilli and Shepard (2007), bias in use-social consequences occur when treatment assigned on the basis of test result vary in quality. A test could be a valid predictor of an outcome but the use of the test might lead to undesirable consequences. It should be noted that this approach requires consideration of factors other than test quality. A fair and unbiased use of test involves more than psychometric validity; it encompasses the consequences to the decision made on the basis of test scores.

Methods of Detecting Test Item Bias in the Measurement of Ability

Many methods of detecting test item bias in the measurement of ability exist and those to be explained in this paper include: item characteristic curve, regression method, chi-square method and transformed item difficulty method among others.

Item characteristic curve approach of detecting test item bias, states that a test is unbiased if all the individuals having the same underlying ability have equal probability of getting the item correct regardless of subgroup membership (Pine, 2006). In other words, an item is said to be unbiased if the characteristic curves for the item measured on two groups are identical. If the situation does not hold, then the item is biased and the area between the group ICCs serve as a measure of the item aberrance (Lord, 2002). All item characteristic curves are plotted from test data and form curves of the same general form: from left to right, beginning low, inclining sharply, and leveling off dramatically as illustrated thus:

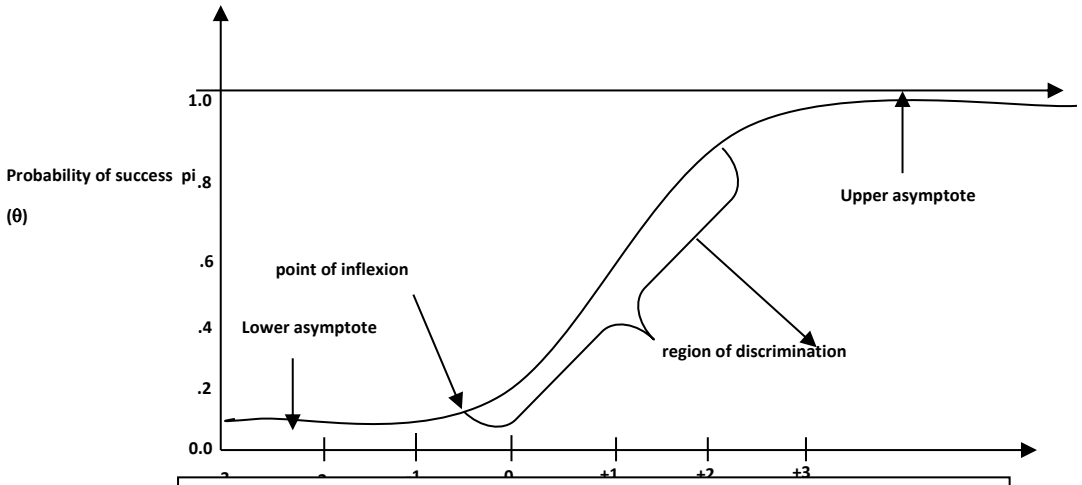

Fig 2: Item characteristic curve for a 3-parametic logistic model

Two things are to be noted in this diagram above and they are: (i) the slope of the curve is monotonic; that is, it always rises and never exactly horizontal, (ii) the two asymptotes, the upper and the lower, which may approach but never, actually reach 1.00 and 0.00 respectively.

Using the regression approach, a test a biased predictor if there is a statistically significant difference between the major and minor groups in the slopes b_{YX} or intercepts K, or in the standard error of estimates SE_Y, of the regression lines of the two groups (Anastasia, 2006). This implies that for a perfectly reliable and unbiased test, the two groups of examinees should share one and the same regression line, and any given test score X should predict the same criterion score Y for a member of either groups with the same probability of error. There ought not to be any systematic under-prediction or over-prediction of criterion performance for persons of either group. If these conditions are not true, then the test is biased if used as a predictor (Joshua, 2005). Using the regression method, it is emphasized that the regression analysis for an item in two different groups should yield equal intercepts, slopes and standard error of estimate in the two groups if the item is unbiased.

In the transformed item difficulty method using major axis, two sets of p-values ought to be computed, one for each population group pair is transformed first, through an inverse normal function Z, and then to delta (δ) values given by δ =4Z+13. From the bivariate scatter plot of the sets of delta values, the absolute values of the perpendicular item plot major distances are used as indications of the magnitudes of item bias. Large deviations show much bias. Sometimes p-values are computed for each group separately and transformed to within group Z-scores using the item mean and standard deviation for that group. The perpendicular line distances are used as indications of the magnitude of the item bias at angle of 45^{O} line.

In the chi-square method, an item is unbiased if for all persons of equal ability, the probability of a correct response is the same regardless of each person's cultural or ethnic group membership. Each major population for comparison is divided into various ability sub-groups on the basis of observed total test scores. Within each score-group, the p-values are computed and compared for the major populations. The expected values for each cell are obtained by multiplying the proportion of examinees who respond correctly to the item within a total score interval by the total number of examinees within the cell. Observed cell values are simply the number of examinees within the cell that respond correctly to the item. For each item, the magnitude of the group difference is indicated by the value of the resultant chi-square statistic divide by its degrees of freedom or mean square (Swaminathan & Rogers, 2005).

Policy Issues in Implementing a Differential Item Functioning Screening Strategy

Clauser and Mazor (2008) posit that the entire domain of item bias is really about policy. In fact, by even considering a bias analysis one is already in the domain of policy. Some organizations have bias analysis legislated whereas others take it on as part of the day-to-day validation process. If bias is being "legislated" from an outside body, according to Zumbo (2009), this legislation will help one to determine the answers to the following policy matters:

1. If there are a lot of different sub-groups to be contrasted, one needs to be clear as to which one are of personal and moral focus. The standard comparisons are based on gender, race, sub-culture, or language.
2. One needs to discuss how much DIF one needs to see before one is willing to consider the item as displaying DIF. In most cases, it is not sufficient to simply rely on the answer that all statistically significant items are displaying DIF because statistical power plays havoc on one's ability to detect effects.

In essence, how much DIF does one need to see before one puts the item under review or study?

3. Should an item only be sent for review if it is identified as favouring the reference group or should an item be sent for review irrespective of whether it favours the reference or focal group?
4. The timing of the DIF analysis is also important. We could consider two scenarios (a) examiner is using a ready-made test; or (b) examiner is developing his own new or modified measure. First of all, in either case, DIF analyses are necessary. In the first scenario where one has a ready-made test that one is adopting for use *and* there is pilot testing planned with a large enough sample, DIF analyses are performed at the pilot testing stage. When one does not have a pilot study planned or one does not have a large enough pilot sample then DIF analyses are conducted before final scoring is done and therefore before scores are reported.

In the second scenario where one is developing a new test, DIF analyses should be conducted at pilot testing and certainly before any norms or cut-off scores are established.
5. What does one do when one concludes that an item is demonstrating DIF? Does one immediately dispense with the item (we won't subscribe to this because the domain being tapped will quickly become too limited) or does one put an item "on ice" until one sends it to content experts and for further validation studies? Part of the answer to this question has to do with the seriousness of a measurement decision.

Implication of Test Item Bias 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 item bias.

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. He concludes that the method of factor analysis holds the most promise. Other tests for unidimensionality include the eigen-value test, the random baseline test, and the biserial test. When factor analysis is used to check for unidimensionality of item, the ratio of first factor variance to second factor variance is used as index of unidimensionality (Hambleton & Cook, 2005). There are possibilities for this assumption to be violated. For example, two items designed to assess multiplication skill in mathematics could be as follows: (i). what is 5×9? (ii). what is the product of five and nine? Item (i) requires only knowledge in mathematical operations, while item (ii) requires for its solution, a specific amount of reading competence as well as knowledge of mathematical operations. When different attributes are being measured as in item (ii), the issue of item bias enters into consideration if such item is administered to two different groups and the responses of one of the groups are dependent on the secondary skill. This type of item measures different types of skills among different groups.

has been observed that some of the national examinations often over-predict or under-predict some candidates from certain states during the selection exercise. Moreover, some examining bodies have different policies of awarding the final grade to examinees. For instance, JAMB has accepted different cut-off points for selecting candidates into Nigerian tertiary institutions based on merit, catchment area, educationally disadvantaged states and institutional discretion. This problem of bias in selection fairness could persist if the examining

bodies do not ensure that examination items have zero error of prediction for all the candidates across the nation, which is a great task to accomplish (Dibu-Ojerinde, 2006). According to him, the language of the test item should be such that it is not apt to be offensive to members of any subgroup. Ideally it is expedient to construct test items in a manner that no element of it would offend examinees on ethnic, sexual, cultural, religious or socio-economic grounds. For instance, a test item on Social Studies which describes, "*Reverend Father Chukwudi Emenike Okafor as the leader of deadly armed robbery gang that has been terrorizing the residents of Makurdi town*" is a biased item. It is obvious that some catholic Christian candidates would be offended because of the implication that '*Reverend Father*' led the *deadly gang* in the first instance. On this term too, Ibo candidates may feel offended that their kinsman is involved in the heinous crime. The candidates from Benue State and those who have relation in Makurdi town may concentrate on the issue in the item instead of the examination. The national examination bodies could take care of bias in such item by reliance on a judgmental approach for detecting and eliminating biased items.

To create bias-free items, the national test developers may ensure that the activities and connotations reflected in the test items are relevant to the life experiences of examinees responding to the items. Test items ought to be written in a straight forward, uncomplicated, easily read manner. Excessive wordiness can obviously interfere with examinees' ability to respond appropriately to test items and therefore constitute bias in the paper. Supposing mathematics item is written like, '*what is the cumulative summation of the integer three when appended to a quantity of an identical nature?'* This is a case of wordiness; the item might be written in a better way as "what *is 3+3*?" or "what *is three plus three?*"

The procedures employed in the administration of the national examination are likely sources of bias. The actual administration of the examinations constitutes a complex interaction among examiners' variables, examinees' variables and situational variables. A situation where candidates in some areas write national examinations in stuffy, poorly furnished, uncomfortable classrooms is an indication that the examinees being tested will perform badly in the examination and sometimes, examination bodies keep mute about this ugly situation, which is of course, an imaginable absurdity. The behaviour displayed by an examiner during examination period can be influential in determining the way that examinees will perform on the test.

Some candidates especially those in the rural areas may be less familiar with the typical testing formats of some examination papers. Such candidates are intimidated by the nature of the test itself. To help in creating uniformity in the national examination, the school teachers and others who are concerned with the performance of the candidates could ensure that examinees are given ample practice opportunities to become accustomed to the formats of the examination items. The classroom teachers ought to see that examinees acquire 'test-wiseness' that may enable them to answer some items correctly through familiarity with testing practices even when they have a limited knowledge of the concept in the items. For instance the longest option in some multiple choice test items may be more likely correct than the short options and this can only be acquired through frequent testing practices (Dibu-Ojerinde, 2006).

Recommendations

It is suggested that vigilance against item bias can help to isolate and expunge biased examining practices in Nigeria. It is therefore desirable that examination bodies use examiners who understand the candidates and the subjects of the examination. The national examination bodies can solve the issue of bias in this situation by appointing qualified examiners who are genuinely committed to the success of the examination generally. It is also important to provide examination administration settings that are conducive in promoting the examinees' best efforts in all the centers throughout the country.

To create bias-free items, the national test developers may ensure that the activities and connotations reflected in the test items are relevant to the life experiences of examinees responding to the items. Examination bodies and educators should take more cognizance of the possibility of test item bias in a testing situation and with this kind of effort, candidates from educationally disadvantaged areas and low socio-economic status would be certain to be fairly treated.

The national examination bodies must try as much as possible to attain unidimentionality in the test and this is one of the very essential assumptions of the test theory and also a vital condition in item bias studies. It will therefore be necessary that the examination bodies should examine the degree of bias in their examinations. In addition, examination bodies should construct test items in such a manner that items are free from writing errors such as wordiness, irrelevancy, offensiveness, and excessive stimulations, so that when an inadequacy exists between groups' examination item scores, the disparity will be attributed to true differences in whatever the test purports to measure in the examinees.

Conclusion

Most societies have now come to conclusion that test results are better criterion for selection purpose than subjective evaluation. Though accepted in most societies as the most objective method of decision-making, nonetheless, the use of test has sparked off some grave concerns among the members of the public in recent years. These concerns have tended to erode people's faith in the power and efficacy of tests. Various criticisms have been leveled against psychological testing and a large proportion of the criticisms pivots on fairness of test to the various identifiable groups taking the same test. It has been affirmed that test bias is inevitably one of the characteristics of examination process which demands a staid attention of examination bodies. Various types of test bias exist and they consist of content bias, atmosphere bias, bias in use-prediction and bias in use-social consequences. Many methods for detecting test item bias in the measurement of ability exist and they include among others: item characteristic curve, regression method, chi-square method and transformed item difficulty method among others. Bias can result in systematic errors that distort the inferences made in any selection and classification.

References

Anastasia, A., & Urbina, S. (2006). *Psychological testing*. New Delhi: Prentice Hall

Berk. (2007). Item Bias detection methods for small samples. Dissertation abstract international.

Camilli, G., & Shepard, L. A. (2009). *Methods for identifying biased test items*. (Vol. 4) Thousand Oaks, CA: Sage Publications.

Clauser, B. E., & Mazor, K. M. (2008). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issues and Practice, 17*, 31-44.

Dibu-Ojerinde, O. O. (2006). Test item bias and implication for testing in Nigeria. *Nigerian Journal of Guidance and Counseling, 3 (1-2), 108-116.*

Emaikwu, S. O. (2011). Evaluation of student's ability in schools. Being a paper presented at a workshop on teaching practice on Friday, 29th July in the College of Agricultural & Science Education, Federal University of Agriculture Makurdi, Benue State

Hambleton, R. K., & Cook, L. (2005). Latent trait models and their use in analysis of educational test data. *Journal of Educational Measurement, 14* (3), 75-96

Joshua, M. T. (2005). Test/item bias in psychological testing: Evidence in Nigerian system. A paper presented at the annual conference of Nigerian Association of Educational Psychologists held at Ahmadu Bello University Zaria from 24th -28th March.

Lord, F. M. (2002). Applications of item response theory to practical testing problems. Hillsdale, New Jessey: Erbaum press.

Lumsden, R.T. (2003). Sampling from a matrix with application to the theory of testing. Princeton University: Statistical research group report no.53

Nunnally, R. S. (2008). Using the Rasch approach to measurement in solving practical school testing problems. *Journal of Educational Measurement, 13* (4), 116 – 124

Pine S. M. (2006). Applications of item response theory to the problem of test bias. Research report of the Department of psychology, University of Minnesota.

Swaminathan, H., & Rogers, H. J. (2005). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 2 (7)*, 361-370.

Zumbo, B. D. (2009). *A Handbook on the theory and methods of differential item functioning (DIF):Scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.