European Journal of Educational and Development Psychology
Vol.4, No.1, pp.42-52, February 2016
Published by European Centre for Research Training and Development UK (www.eajournals.org)

DETECTING ITEM BIAS IN RIVERS STATE JSSCE BUSINESS STUDIES USING ITEM RESPONSE THEORY (IRT) APPROACH

Dr. (Mrs) Goodness Orluwene and Asiegbu Chinelo Nneka

Department of Educational Psychology, Guidance and Counselling, Faculty of Education, University of Port Harcourt

ABSTRACT: *This study identified biased test items in the Rivers state JSSCE Business Studies conducted in 2009 using Item Response Theory Approach. The Sample comprises 4000 JSS III students in public and private schools that sat for the JSSCE Business Studies Examination in 2009 in Rivers state which were selected through stratified proportionate random sampling. The instruments for data collection were the test items in JSSCE Business Studies conducted in 2009 and a result collection sheet. Two null hypotheses drawn from two research questions were tested at 0.50 probability of correct response. The assumptions of unidimensionality and local independence were checked using factor analysis. The students' responses to the test items were calibrated with an IRT statistical software named ex-calibre 4.2.2 developed by Assessment Systems Corporation. The results revealed that the test items met the assumptions of unidimensionality and local independence. It also revealed that there is school type bias in the Rivers State JSSCE Business Studies test items for 2009. Based on the findings, recommendations were made, one of them being that business studies curriculum should be restructured so as to accommodate the new technological advances which may be one of the reasons for the presence of school type bias in the Business Studies test conducted in 2009 in Rivers State.*

KEYWORDS: Item Response Theory, Item Bias, Item Characteristic Curve, Unidimensionality

INTRODUCTION

The assessment of learning outcomes by using tests is one of the basic issues in education. Test results help educators know how students learn and provide feedback for restructuring and modifying the teaching – learning process. Tests are also often used to make important decisions about people. In educational system, institutions use scores from tests to make decisions on admissions like placement of students in gifted schools. Some companies also make job decisions about people based on test scores. There is therefore the need to ensure that tests administered by different examination bodies are as fair as possible for all subgroups of a population.

A fair test is one in which different sub – groups of a population with equal ability have the same probability of correct responses to the test items. According to Roover cited in Perrone (2006), "a fair test is one that is comparably valid for all groups of individuals and that affords all examinees' equal opportunity to demonstrate the skills and knowledge which they have acquired and which are relevant to the test's purpose". As test results are often the basis for decision that affects student's educational future, tests should provide the same and equal opportunities for all students to demonstrate their skills, abilities and knowledge. It therefore becomes necessary to avoid bias which may negatively influence examinees' scores.

Bias then, according to Hambleton & Rogers (2012) is the presence of some characteristics of an item that result in differential performance for individuals of the same ability but from different ethnic, sex, cultural or religious groups. This means that an item is considered biased when different groups but with the same ability level, have nevertheless, different probabilities of answering the item correctly. Biased items can lead to biased measurement of ability because the measurement is affected by so-called nuisance factors (Ankerman, 2002).

The presence of bias in test items threatens the content validity of such tests. This means that the items are measuring attributes that are not necessary or relevant to the construct being measured by the test. A valid and reliable examination indicates that the exam has been carried out in accordance with its target and test items were prepared in line with the purpose of the exam. If the item in the test provides an advantage to any of the group taking the test because of various features like school type, it can be said that the test has bias in favour of that group. This will negatively affect the validity of the decisions made based on the test scores. Bias can also lead to systematic error which may, according to Zumbo (2012), distort the inferences made in the classification and selection of students.

Business Studies is an arm of vocational education which prepares the youth for future membership and participation in the life of the society and for its maintenance, growth and development. It is meant for skill acquisition and is designed to prepare the students for the world of work. At the Junior Secondary level, it has five arms which include commerce, office practice, typing, shorthand and book keeping.

Multiple choice tests are most commonly used in tests for measuring achievement and cognitive ability. Tests like the JSSCE Business Studies are usually used to measure how well students' understand Business Education at the junior secondary level and for promotion to the next class. The negative effect of biased test items in Business Studies can hinder academic achievement by not seeing and taking into consideration individual differences in the way different groups like public and school students learn. It therefore becomes very necessary to avoid bias in the test items which may unfairly influence examinees' scores as per the group they belong to (Hambleton & Swaminathan 1985).

A biased Business Studies Test may be due to the presence of irrelevant constructs related to school type and daily life experiences. It has also been observed that most questions in the Rivers State JSSCE Business Studies are often repeated yearly. If those items are biased and are not screened, the validity of the test items will be jeopardized. More so, in the analysis of the JSSCE Business Studies results, school type is not equally represented. Discrepancies have been observed in the result in favour of one sub-group of the population or the other. Therefore, it is necessary to ensure that Business Studies test do not contain items that will function differentially in different sub-groups of the population and which may cause bias to students with regards to gender and location.

Item response theory provides a unified framework for conceptualizing and investigating bias at the item level.Test developers and educational researchers have developed a number of item bias detection procedures. Ertuby & Russele (1999), Ojerinde (2012) and Osterlind (2014) suggested that because of their high level of sophistication, IRT, procedures provide the best results for detecting biased items. This means that biased items in Rivers State JSSCE Business Studies can be detected using an IRT approach.

Mellenberg (1994) posited that "IRT is a model for expressing the association between an individual's response to an item and the underlying latent variable often called ability or trait being measured by the instrument. IRT models use item responses to obtain scaled estimations of θ as well as to calibrate items and examine their properties". From Psychology Wiki (2012), "IRT is a body of theory describing the application of mathematical models to data from questionnaires and tests as a basis for measuring abilities, attitudes or other variables". It applies mathematical functions that specify the probability of discrete outcome such as correct response to an item in terms of person and item parameters. Wikipedia foundation (2012) has a similar definition. It sees IRT as "a paradigm for the design, analysis and scoring of tasks, questionnaires and similar instruments measuring abilities, attitudes or other variables". In agreement to this, Lord (1980) states that the main task in most testing work is to infer to the examinees ability level or skill. In order to do this, something must have to be known about how an examinees knowledge or skill determines his response to an item. Thus IRT starts with a mathematical statement of how response depends on the level of ability or skill. Osterlind (2014) believes that it is a collection of models that provide information about the properties of items and scales they comprise of through analysis of individual item responses . IRT is also called a latent theory because the theory assumes the existence of a latent trait which is a tester characteristic that leads to a consistent performance in a test.

Reeve (2003) believes that IRT refers to a set of mathematical models that describe in probabilistic terms the reason between a person's response to an item and his\her level of the latent variable being measured. In order words, IRT assumes that there is an underlying probability that someone with a particular level on the latent trait will endorse a given item. The latent variable is usually a hypothetical construct, trait, domain or ability which is assumed to exist but cannot be directly measured by a single observable variable but can be measured indirectly using multiple items in a multi item scale (Wikipedia Foundation 2012). The person's level on the construct is assumed to be the only factor that accounts for his or her responses to each item in a scale. For example, a person with a high ability level in Business Studies will have a high likelihood of responding correctly most of the time to items in a Business Studies achievement test, whereas a person with a low ability level will have a low likelihood of responding correctly except for guessing. This means that IRT depicts a more thorough picture of item functioning as it predicts how an examinee will perform on a test item. This, according to Crocker & Angina (1986), Zhu (2006) allows for comparisons to be made between examinees that have taken different tests and also allows for bias detection.

IRT methods have very great capabilities for diagnosing measurement problems. A test developer using an IRT technique generates a mathematical function to describe the relationship between test performance and ability or trait (Reid, Kolakowsky, Lewis & Armstrong, 2007). If a test has been developed using IRT, the level of the ability and trait measured can be estimated for an individual completing any subset of the test items. Administration of the entire test may no longer be necessary.

IRT also has potent tools for studying potential bias in assessment instruments. It is from IRT that the item characteristics curve (ICC) approaches to the detection and correction of test item bias are derived (Osterlind, 1983). It has been seen as particularly useful for studies of item bias because in theory, the ICC does not depend on the distribution of ability in the sample used to ascertain the parameters. Hence, if the parameters of an ICC are estimated separately for two samples drawn from the same population, the resultant curves are supposed to be the

same even though the sample may differ in the distribution of abilities within them. If the curves are not the same, the conclusion of the bias may be necessary.

IRT is a mathematical and statistical model of item responses in a population of individuals (Birnbaum, in Gregory 2004). It is founded on two basic ideas, one is that an individuals' ability, knowledge and skill can be predicted. Secondly, the relationship between an individuals' knowledge, skill and ability and the responses to the test items in a test can be described by an Item Characteristics Curve (ICC). Objectively, IRT according to Hambleton (1993) "aims at looking at an observable trait (an individual's performance) and an unobservable trait (an individual's ability) on a test which in turn will be measured by a mathematical function". Harris (2012) believes that its goal is to model a relationship between a type of variable that may not be observed as typified by an individuals' ability and the likelihood that that individual may get the correct answer.

A variety of models have been developed from the IRT perspective and these models differ from each other in at least two important ways. One important difference among the measurement techniques is in terms of the item characteristics or parameters that are included in the model. A second important difference is in the nature or type of the response option format. For test items that are dichotomously scored, their IRT models are known as one, two and three parameter models, 1pl, 2pl and 3pl for short.

The simplest IRT model is the 1pl which contains only a single parameter and is usually described as the Rasch model. For the Rasch Model, items are believed to vary in their relative difficulty. Once the relative difficulty of each item is estimated, person estimates of test scores can be obtained by subsequently treating the item difficulty as fixed.

The two parameter model adds a varying slope to the Rash model allowing item responses and person abilities to interact. In this model, each item varies in average reliability. Using the 2pl as a baseline, the 3pl model further adds a "pseudo-guessing" parameter with the intent of accounting for observed performance of these persons with very low levels of the latent trait.The 2pl uses both difficulty and discrimination parameters. The difficulty parameter (b) tells us how easy or how difficult an item is. It is the only parameter used in the 1pl model. The discrimination parameter is also known as the (a) parameter and it shows how efficiently an item can distinguish between highly knowledgeable students and less knowledgeable students. The 3pl uses a, b and c parameters. The c parameter is the guessing parameter and its value shows how likely the examinees are to obtain the correct answers by guessing. When the guessing parameter is taken into consideration, it portrays the fact that in some items, though the examinee may not know anything about the subject matter, he or she can still have a probability of getting the right answer.

The models discussed above are designed to be used for binary outcomes as the response option, for example, correct or incorrect. However, many tests, questionnaires and inventories include more than two response options like strongly agree, agree, disagree, strongly disagree. Such items are known as polytomous items and they require IRT models that are different from those required by binary or dichotomously scored items. They are called polytomous IRT models. Examples are the Graded Response Model and the Partial Credit Model.

Item characteristic curve (ICC) is the functional relationship between the proportion of correct response to an item and a criterion variable (Baker & Kim, 2004). Psychometricians use ICC to describe and evaluate characteristics of the items in a test. It reflects the probability with which

the individuals across the range of trait levels are likely to answer each item correctly. It models the relationship between a person's probability for endorsing an item category and the level of construct measured by the scale.

Osterlind (1983) sees it as a graphic presentation of a mathematical function that describes the probability of an examinee correctly answering a test item relating to the ability measured by the total set of items of the test. ICC shows the relationship between the individuals' ability on an item and the likelihood that the individual will respond correctly to that particular item (Suen, 2010). ICC is assumed by researchers to be the major concept of IRT as it is a monotonically increasing function which relates the relationship between examinees' item performance p(Θ) and the trait (Θ) underlining item performance. The regression of the score of examinee's ability, otherwise known as the Item Response Function (IRF) is shown in Figure 1.

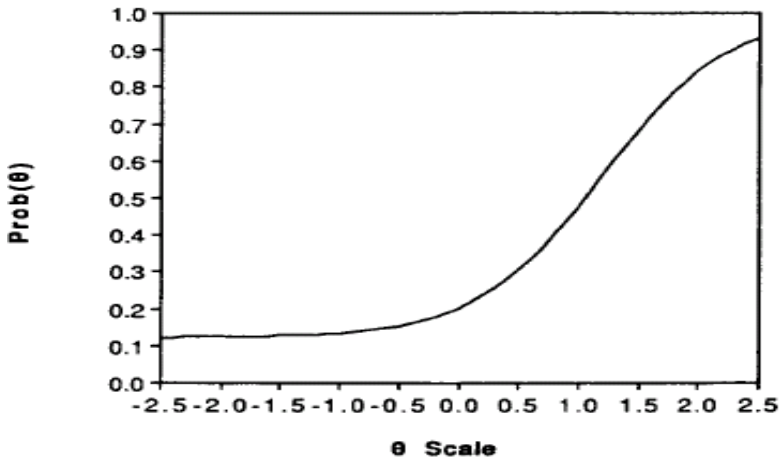


Figure 1: Item Characteristics Curve (ICC) from Harris (2012)

On the x-axis is the function together with the examinee's ability level while the probability to answer an item correctly is shown on the y-axis. Basically, every candidate or examinee is presumed to acquire some level of ability (represented by Θ) and is thus placed on the ability scale according to that level. For every examinee, there is a probability that he/she, according to his/her ability, will be able to answer an item correctly. This probability, represented by P(Θ), is lower for individuals with lower ability levels and higher for those with higher levels in a given situation. Thus, a plot of probability function P(Θ) against (Θ) will produce an S-curve, which depicts the form of the ICC.(Baker, 2012, Lord, 1980, Warm, 2008). This means that if the probability function of students' Achievement test in JSSCE Business Studies is plotted against their ability levels, a typical S-shaped form of the ICC (Figure 1) is supposed to appear. Each item in the test will have a separate ICC, the shape of which will be determined by the type of ICC model used.

IRT has caused significant changes in psychometric theory and test development. As noted by Hambleton & Slater in Marie (2004), it basically assumes that a single trait depicts an examinee's performance on a given test and that the probability of a correct response on an item is a monotonically ascending curve. From IRT, Differential Item Functioning (DIF) is derived which is a powerful method for investigating item bias. The ICC of an item shows graphically the probability of the correct response as a function of the magnitude or level of the

underlying trait being measured. Osterlind (1983) describes ICCs as "the most elegant of all the models to tease out item bias".

The purpose of this study therefore is to detect items showing bias in JSSCE Business Studies conducted in Rivers State in 2009 with regards to *school type* using IRT approach. The following research questions guided this study:

1 To what extent do test items in Rivers State JSSCE Business Studies conducted in 2009 comply with the assumptions of unidimensionality and local independence?
2 To what extent do test items in Rivers State JSSCE Business studies conducted in 2009 show up school type differences?

The following null hypotheses were tested:

1. The test items in Rivers State JSSCE Business Studies conducted in 2009 will not differ significantly in complying with the assumptions unidimensionality and local independence.
2. The ICCs of students' responses to the test items in Rivers State JSSCE Business Studies conducted in 2009 will not differ significantly due to school type at 0.50 probability of correct response.

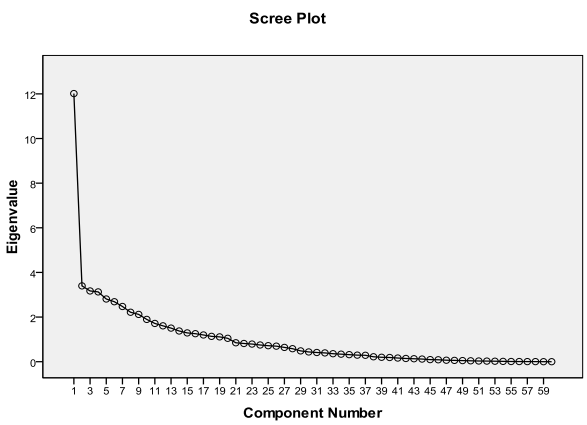
METHOD

Analytical descriptive research design was adopted for this study because it involves the in depth study and evaluation of available information in an attempt to explain a phenomenon. It is primarily concerned with interpreting relationship by analyzing facts or information already available. The population comprised 52893 students in private and public schools that sat for the JSSCE Business Studies conducted in Rivers State in 2009. Proportionate stratified random sampling was used to obtain a total of 4000 students consisting of 2821 in public schools and 1179 students in private schools. Data were extracted from the students' responses to the test items with the aid of staff members, exams and records department, Rivers State Ministry of Education with a result collection sheet. Factor analysis with the aid of SPSS was performed on the students' responses to the test items so as to check if the data met the assumptions of unidimensionality and local independence. An IRT software, Ex- caliber 4.2.2 developed by Assessment Systems Corporation was used to estimate the item parameters with the marginal maximum likelihood estimation technique and to also draw the ICCs.

RESULTS:

Hypothesis 1:

Fig. 2:

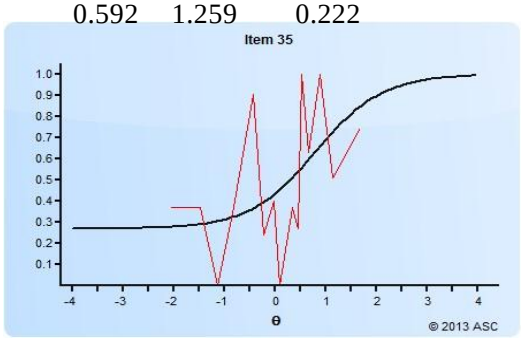


From fig. 2 above, the hypotheses that the test items in Rivers State JSSCE Business Studies conducted in 2009 will not differ significantly in complying with the assumptions of unidimentionality and local independent therefore upholds. This is because the eigen value of the first factor in figures 2 is large compared to the second factor and the eigen value of all the other factors are all about the same.

Hypotheses 2: Fig 3 below (1 – 8) shows the ICCs of the items that are not biased for the public and private school students that sat for the JSSCE Business Studies in 2009

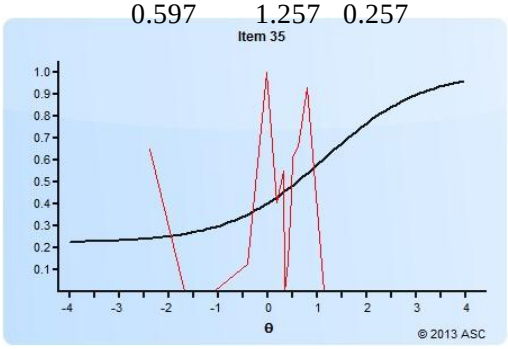
Public

a	b	c
0.592	1.259	0.222



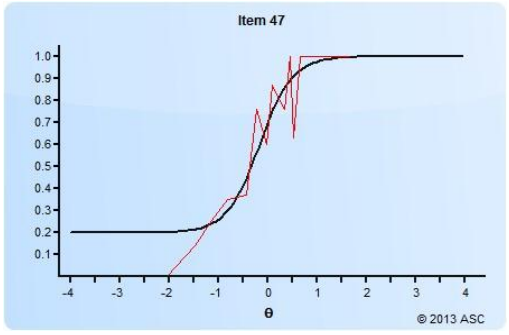
Private

a	b	c
0.597	1.257	0.257



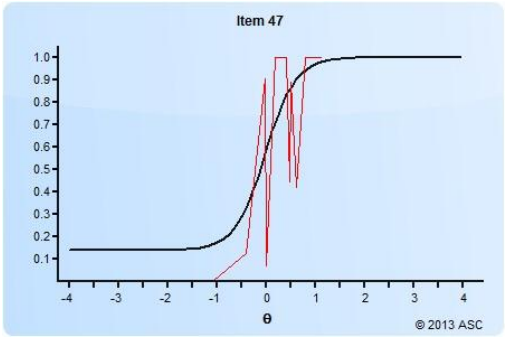
Public

a	b	c
1.878	0.056	0.198



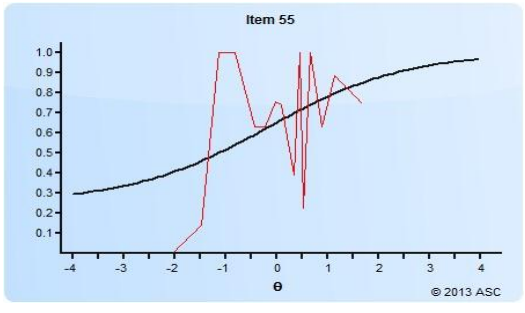
Private

a	b	c
1.873	0.051	0.139



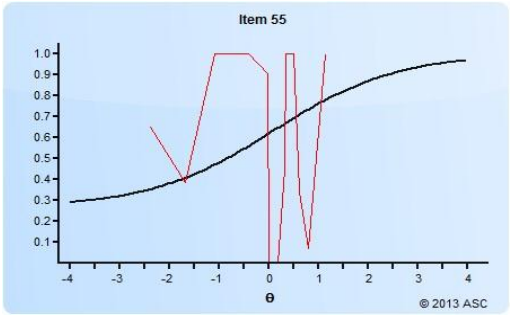
Public

a	b	c
0.429	0.168	0.253



Private

a	b	c
0.427	0.168	0.266



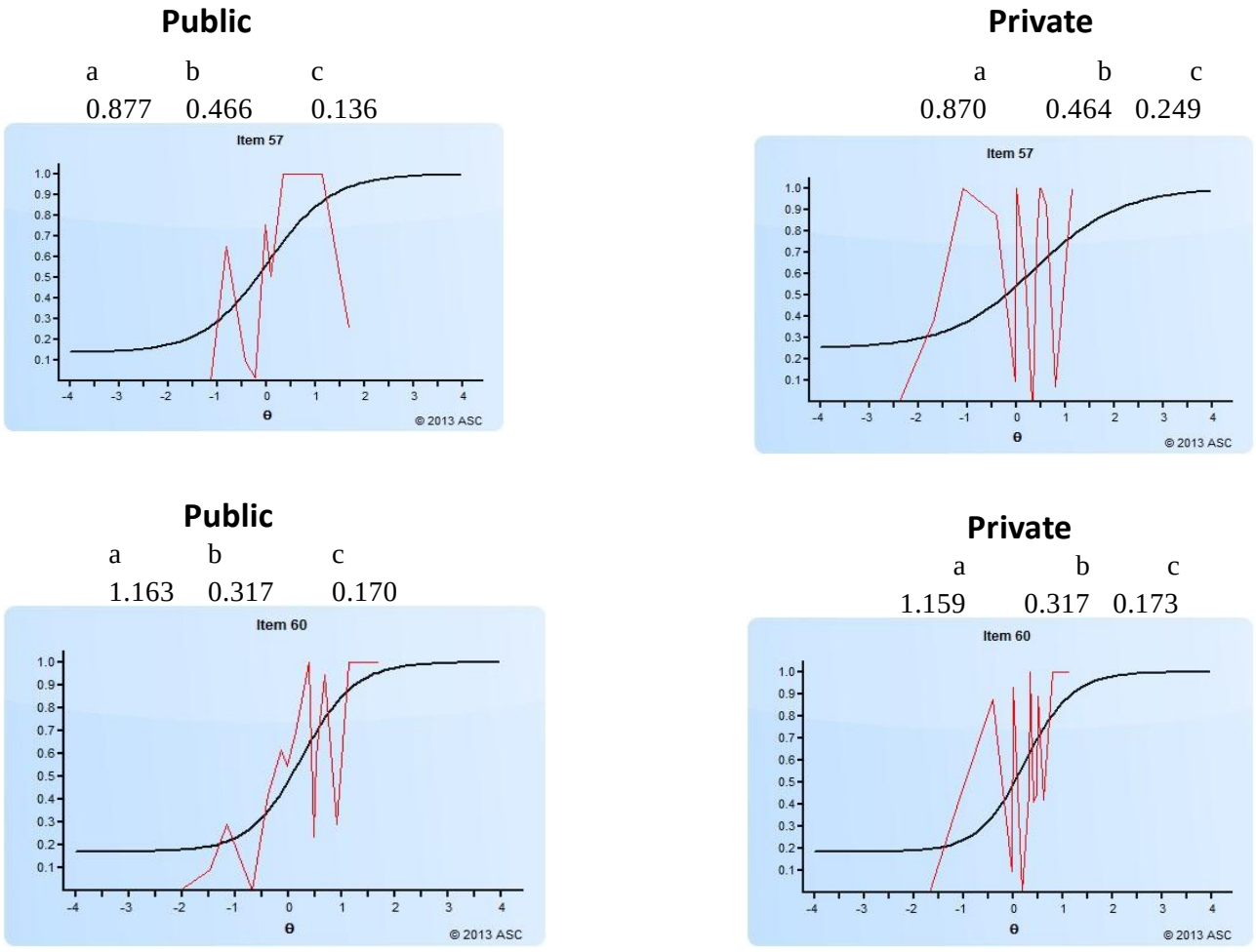


Fig. 3: (1-5)

Only 5 ICCs out of 60 are similar for students in public and private schools that sat for the JSSCE Business Studies conducted in 2009.

DISCUSSION

The calibration result shows that out of 60 multiple choice test items, 5 items were equally difficult for the two groups. They are items 35, 47, 55, 57, and 60. There was no variance in two items which are items 1 and 59. Twenty-four items (2, 3, 7, 10, 11, 12, 13, 14, 15, 20, 21, 23, 24, 28, 30, 32, 37, 39, 40, 44, 45, 51, 52, and 58) were biased against students in private schools while 29 items (4, 5, 6, 8, 9, 16, 17, 18, 19, 22, 25, 26, 27, 29, 31, 33, 34, 36, 38, 41,42, 43,46, 48, 49, 50, 53, 54 and 56) were biased against students in public schools. This agreed with Uremu & Onwuka (2013) whose result findings shows that of all the items flagged as biased due to school type, greater number were in favour of students in private schools. They opined that private schools pay more attention to their students and jealously guard their equipment. They also go out of their way to prepare their students for external examinations by organizing after-school, week-end and holiday lessons so as to enhance their students' performance and so boost the image of their schools.

However, a closer look at the items reveals that students in public schools endorsed items related to typewriting and accounting correctly more than those in private schools. This may be because of the emergence of computers where most schools no longer use typewriters to teach, though subject experts will still set questions based on typewriter usage. It could also be due to lack of qualified teachers in private schools that can handle such subjects as most of the privately owned schools find it difficult or are unwilling to employ high-profile, qualified and competent teachers that attract huge fund. Research findings of Suobene (2008) in agreement to this reveals a dearth of qualified technical and vocational subject teachers in private schools. It also reveals that most private schools do not employ qualified teachers so as to reduce the cost of managing the schools. The reverse is the case in public schools where funding of schools, recruitment of teachers and provision of equipment are solely the responsibilities of the government. Item 1 was extremely easy for the students most especially those in public schools and therefore calls for re-phrasing, rewording or re-structuring.

Item 59 was also very difficult for both groups. This might be because there was a mistake in the answer options which therefore requires correction. Items 25, 29, and 33 were also difficult for both groups. Students found item 25 very unfamiliar. They might not have seen any other type of pen apart from the usual one they use in writing. They might equally not have seen a stencil because they are quickly fading away. Item 29 had a long sentence structure that must have confused the students while item 35 is controversial because most schools no longer use typewriter but rather computer to type.

Finally, all the other items did not have any clue in the structure or content that will make them function differentially, so there might be other factors other than item bias that made the ICC's to be different which is therefore subject to further investigation.

CONTRIBUTIONS, CONCLUSION AND RECOMMENDATIONS

The findings from this study reveal that there is school type bias in the Rivers State JSSCE Business Studies conducted in 2009. More items were biased against students in public schools. It was also discovered that students in public schools excelled in items that were based on typewriting and book keeping or accounting. On the basis of these findings, the following implications are deduced:

1. JSSCE Business Studies conducted in 2009 favours one sub group more than the other. As some of the items are repeated yearly. It therefore calls for revision, rephrasing or rewording of the faulty items.

2. Some aspects of the Business Studies curriculum should be restructured or revised.

3. There may be other factors outside bias affecting students' performance in Business Studies as some of the test items do not have school type clues.

The researchers therefore recommend that test developers in the state and other examination bodies should ensure that their items are bias free by utilizing IRT approach, most especially the Item Characteristic Curve in teasing out biased test items which will subsequently be rephrased or removed entirely from the item bank. Secondly, Business Studies curriculum should be restructured so as to accommodate the new technological advances which may be one of the reasons for school type bias in the business studies test conducted in 2009.

REFERENCES

Adedoyin, O.O. (2010): Using IRT Approach to Detect Gender Biased Items in Public Examination: A case Study of the Botswana Junior Certificate Examination in Mathematics. *Educational Research and Reviews,* Vol. 5 no. 7. pg 385-399.

Ankerman, T.A (2002): Practical Applications of Item Response Theory, sponsored by the Department of Measurement, Statistics and Evaluation Seminar and Training Center at the University of Maryland. College Park.

Baker, F.B. (2012). *The Basics of Item Response Theory. Eric Cleaning house on assessment and Evaluation*. University of Maryland and College part. Retrieved on 6/06/2012 from *http;//edres.orggirl./baker/.*

Baker, F.B., & Kim, Seock-Ho (2004*): Item Response Theory, parameter Estimation Techniques*. New York. Marcel Decker, Inc

Crooker, L. & Angina, J. (1980): *Introduction to Classical and Modern Test Theory.* New York: Holt Rinehart & Winston.

Ertuby, C. & Russell, R.J.H. (1996): *Dealing with Compatibility Problems of cross Cultural Data.* Paper presented at the 26th International Congress of Psychology. Montreal. August 1996.

Gregory, R.J. (2004): *Psychological Testing, Principles and Applications*. Boston. Allyn and Bacon.

Hambleton, R. & Rogers, J. (2012): *Item Bias Review.* Practical Assessment, Research, and Evaluation. Retrieved on 12/06/2012 from *http://PAIEonline.nrt/geton.aspv=48 n=six.*

Hambleton, R.K (1993): *Principles and Selected Applications of Item Response Theory.* In R.L Linn (Ed). Educational Measurement (3rd ed). Phoenix Arizona. American Council on Education and the Oryx press.

Hambleton, R.K. and Swaminathan, H. (1985) *Item Response Theory, Principles and Applications*. Boston. Kluwer-Nijhof Publishing co.

Harris, D. (2012); Comparison of One, Two, Three, parameter IRT Models. The Instructional topics in Educational Measurement Series. Pg 153-163. Retrieved on 06/06/2012 from *http://www.ncme.org/pubs/items.cfm*

Lord, F.M. (1980); *Applications of Item Response Theory to Practical Testing Problems*. Mahwah, New Jersey. Lawrence Erlbaum Associates.

Marie, D. B. (2004): Use of Differential Items Functioning (DIF) analysis for Bias Analysis in Test construction. *SA Journal of Industrial Psychology*.Vol .30, No. 4 p. 52-58.

Mellenberg, G.J. (1994): A One-dimensional Trait Model for Continuous Item Response. Multivariate Behavioral Research. 29. pg 223-236.

Ojerinde, D.(2012): *Introduction to Item Response Theory, Parameter Models, Estimation and Application*. Abuja- Nigeria, Marvellous Press.

Osterlind, S. B. (2014): *Constructing Test Items*. Boston. Kluwer Academic Publisher.

Osterlind, S.J (1983); *Test Item Bias*. Newbury Park, London. Sage Publications.

Pedrajita, J.O. & Talisayon, M.Y. (2009): Identifying Biased Test Items by Differential Item functioning Analysis using Contingency Table Approaches. *Education Quarterly*, Vol. 67 no 1 pg 21-43.

Reeve, B.B. (2003): Item Response Theory Modeling in Health Outcomes and Measurement. *Expert Review of Pharmacoeconomics and Outcomes Research*, Vol. 3 no. 2. Pg 131-143.

Suen, H.K. (2010): *Principles of Test Theories*. Hillsdale. N.J. Lawrence Erlbaum.

Suobere, T.P. (2008): Constraints too the Effective Implementation of Vocational Education Program in Private Secondary School in Port Harcourt Local Government Area. *Asian Pacific Journal of Co-operative Education*. Vol. 9 No. 1 p. 59-71.

Uremu, O. & Adams, O (2013): Differential Items Functioning Method as an Item Bias Indication. *Journal of Educational Research*. Vol. 4 No 4. p. 367-373.

Warm, T.A. (2008): *A Primer of Item Response Theory (Technical Report*). Oklahoma City. US Coast Guard Institute.

Zhu, W. (2006): Constructing Test Items Using Item Response Theory in Wood T. and Zhu W, *Measurement Theory and Practice in Kinesiology.* Champaign. Illinois Human Kinetics.

Zumbo, D. B. (2012): A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-type Item Scores. Retrieved on 06/06/2012 from: *http://edu.ubc.ca/faculty/zumbo/dif/handbook.pdf*.