Kashere Journal of Education 2022, 3(1): 146-158 **ISSN: 2756-6021 (print) 2756-6013 (online)**
DOI: https://dx.doi.org/10.4314/kje.v3i1.19 **Creative Commons Attribution License (CC BY 4.0)**
Aituariagbon, K. E. and Osarumwense, H. J. **©2022 Federal University of Kashere**

Non-parametric Method of Detecting Differential Item Functioning in Senior School Certificate Examination (SSCE) 2019 Economics Multiple Choice Items

Aituariagbon, K. Ehis and Osarumwense, H. Judith
Department of Educational Evaluation and Counselling Psychology, Faculty of Education, University of Benin.
kinsb2003@gmail.com 07039307359
hannahujudith@yahoo.com, hannah.osarumwense@uniben.edu 08025992106

Abstract
This study analyzed Non-Parametric Methods of Detecting Differential Item Functioning (DIF) in SSCE 2019 Economics Multiple Choice Items. Three (3) research questions were raised to guide the study. The survey research design was employed. The population of the study consists of 14,400 SS II Economics students from Edo South Senatorial District of Edo State, Nigeria. The number of sample size used in the study was 1,440 SS II Economics students. The stratified random sampling technique was adopted for this study. The instrument used to gather information was 60 Economics multiple choice items of NECO SSCE 2019 June/July. The instrument was validated by experts in Economics and Measurement and Evaluation and Quality Assurance Unit of National Examination Council (NECO) officials. Reliability was determined using Cronbach's Alpha Statistics and coefficient of 0.72 was obtained to detect the items that functioned differentially by sex, three (3) DIF methods (Mantel Haenszel, Standardized p-diff and transform item difficulty (delta plot) were used. The findings showed that the three (3) methods displayed DIF. It was found that standardized p-difference and transform item difficulty are most suitable methods than Mantel Haenszel statistics in detecting 2019 Economics Multiple Choice Items prepared by SSCE 2019. It was also recommended that other non-parametric statistics be used to detect differential item functioning.
Keywords: Differential Item Functioning, Mantel Haenszel, Transform Item Difficulty and Standardized P-Difference.

Citation: Aituariagbon, K. E. and Osarumwense, H. J. (2022). Non-parametric Method of Detecting Differential Item Functioning in Senior School Certificate Examination (SSCE) 2019 Economics Multiple Choice Items. *Kashere Journal of Education,* 3(1): 146-158.
Submitted: 15/2/2022 **Accepted**: 17/4/2022 **Published**: 1/6/2022

Introduction

Testing in education and psychology is an attempt to measure examinee's knowledge, intelligence or other characteristics in a systematic approach. Test is an assessment intended to measure the student's knowledge, skill, aptitude, physical fitness. Therefore, the importance of testing in education system cannot be overemphasized. The purpose of testing is to reveal the level of the latent trait of a testee. There is no doubt that testing has become one of the most important parameters by which the society adjudges the product of their education system (Emaikwu, 2012). The accuracy and consistency is important in the content of test which emphasized the use of reliability and validity. If a test is reliable, it is assumed that the test can be extended to a population and to a variety of condition. If a test measures what it ought to measure, it can be thought of as the validity of the measurement (Adediwura, 2013). Tests are said to be fair, if the score interpretations are valid for all relevant subgroups. The factors that influence the magnitude of validity and fairness of examinee achievement test are the clear definition of constructs (skills and abilities) that are intended to be measured, items and tasks that are explicitly designed to assess the construct.

A fair test is not expected to discriminate against sub-groups of examinees or give an

advantage to other groups. When test contents enable all examinees to demonstrate what they know and what they can do irrespective of their language, culture, school ownership, school location and sex. A good and ideal test is expected to measure student's ability accurately and include some characteristics such as unbiased and free of error, constituently and entirely own construct, content validity and obtained an adequate sampling of the examinee's domain of learning. Therefore, it would be applied appropriately; specific aims can be easily and accurately interpreted. It is not an easy task to achieve test fairness particularly in a real world situation because examinees could be from a different language, ethnic, school ownership and culture. Testing various performances of examinees could be based on local language distant from the target language or it could be based on unfamiliar format or emphasis on a learning domain (affective, cognitive and psycho-motor) that is not stressed in the education system of the examinees school or location (Yaghoobi, 2016).

Therefore, it is unrealistic for a test to be fair. It can only be equitable, hence, the reason why examinees respond to an item differently. Test fairness in the content of test items does not prevent items from functioning differentially for different sub-groups. Moreso, differential items functioning can be regarded as a statistical difference between the probability of a specific population group getting the item right and a comparison population group getting the item wrong given that both groups have equal level of expertise with respect to the content being tested. DIF occurs when members of a particular group receive more vigorous coursework over other group or when examinees of a group attend better schools than the examinee of the other group.

Zumbo, (2007), opined that learners who have similar knowledge of the test materials on a test based on total examination result should perform similarly on individual examination irrespective of Sex, culture, ethnicity and race. Zumbo (2007) explained the reasons why an item can be bias: A biased item measures attributes irrelevant to the tested construct, more often, examination items are considerable bias because they contain sources of difficulty that are not relevant to the construct being measured and these extraneous sources impact test taken performance.

Differential Item Functioning (DIF) is presents when examinees of approximately equal knowledge and skills in different groups perform in substantially different ways on a test question. The role of DIF is to identify items that could be unfair, because group difference in relevant skills and abilities have been taken into consideration. Also, DIF is a cause of great concern, considering the examinee`s result which is taken to be good indicators of student`s ability level of performance, (Ndifon, Umoinyang & Idiku, 2010). DIF is also regarded as "measurement bias" when testees from different groups with same latent trait have different probability of giving a certain response on a test. An item cannot flag DIF, if examinees from different groups have different probability to give a certain response; an item only flag DIF if and only when examinees from different groups with the same underlying true ability have different probability of giving a certain response.

Differential item functioning occurs if respondents to a questionnaire or test item from different groups with the same overall scores have different probabilities of giving a correct or positive response to the item. Specifically, DIF has been recognized as a standard tool of measuring significant item function difference across groups (such as Sex or race) while controlling the overall scores on the trait being measured (Zhang, 2015). Differential item functioning occurs when examinees of different groups show differing probabilities of success on the item after matching on the construct that the item is intended to measure. Differential item functioning refers to differentiations between the correct answering probabilities of examinees in different groups to the related item in a comparison to be made on ability level which the item intends to

measure (Zumbo, 2007). Differential item functioning helps to spot item that may be unfair but DIF is not synonymous with bias. The major reason DIF is not a proof of bias is because the matching process is imperfect. Test usually measures a single trait, skill or knowledge. A fair item or question may show DIF when measuring a skill that is not well represented in the test as a whole.

Acar and Kelecioglu (2010) viewed DIF as the differentiation between the item and the probability of correct response to the item in every latent trait level of psychological structure that will be measured. DIF occurs when group differences in performance of a test item (Buzick & Stone 2011). Test items are identified as exhibiting differential item functioning when, after matching an examinee group by a measure of ability, the performance of one group is significantly higher than the other group, on average. When DIF occurs, it means that a test item measures traits or abilities that are secondary to the target ability. For Economics students, the trait could be, to assess the test-taken on the statistical aspect of Economics using the appropriate formular to respond to the items. Ogbebor and Onuka (2013) stated that differential item functioning is an approach that is widely used to find out item that are bias. In other words, DIF is a tool, concept or an instrument that can be used to discover item that are bias in a test. Differential item functioning is a statistical technique used to assess the existence or the presence of item bias. It is a systematic error in the predictive or constructs validity of an item that may be attributed to factors irrelevant to the test. Research has showed that the modern approach for detecting item bias is by providing evidence of differential item functioning.

French, Hand, Nam, Yen and Vazquez (2014) explained the concept of DIF as a situation where examinees from two groups who have equal levels on the measured ability have different probabilities of endorsing the same item response. Moreso, the presence of DIF shows that the item is not performing the same across groups. Locating items on which group of examinees perform significantly better than another group is logically the first step in detecting item bias. In IRT context, if the items exhibit DIF, then the ICCs will be identified differently for the groups. The ICCs can be identifiable in two common ways. First, the curve can differ only in threshold. That is, difficulty parameter and hence, the curve are displaced by the shift in their location on the theta continuum of variation. Second, the ICCs may differ not only on difficulty but also on discrimination and or guessing and hence the curve may be seen to intersect. However, this study is not on parametric approach which used the ICC (IRT). The key decision that must be made for DIF analysis is selecting the appropriate model, hence, the need for IRT model and non IRT model in non-parametric (classical test theory). Different models allow a different number of item parameters (i.e., b, a, c parameters) to be estimated from the data of item responses and thus, allow for evaluation of DIF for different item properties.

Non-Parametric Methods of Detecting DIF

Mantel Haenszel Method of Detecting DIF

Mantel Haenszel DIF procedure developed from Mantel and Haenszel (1959) but was proposed as a method for detecting DIF by Halland and Thayer (1988) Applying the M-H method DIF detection is done by grouping examinees according to an estimate of ability (which is the total test score) and then forming a two by two (2x2) contingency table crossing group membership (reference and focal) and item performance (correct and incorrect) for each level of ability.

Gierl, Jodoing and Ackenman (2000) explained that the M-H statistics is distributed as a chi-square test with a degree of freedom usually one. It generally stated that in its null hypothesis, there is no relationship between group member and test performance on one item after controlling the ability. He further stressed that, M-H is used to estimate the constant odd ratios that

Kashere Journal of Education 2022, 3(1): 146-158 **ISSN: 2756-6021 (print) 2756-6013 (online)**
DOI: https://dx.doi.org/10.4314/kje.v3i1.19 Creative Commons Attribution License (CC BY 4.0)

yield a measure of effect size for evaluating the amount of DIF that is present.

Interpreting the M-H statistics in DIF, if the MH chi-square statistics is significant, the item is considered to be performing differentially for one of the compared groups. Also, if the difference measured is greater than one, the item is performing differentially in favour of the reference group; if it is less than one, then it is performing differentially in favour of the focal group. Contingency table for an item for reference and focal group with test score k.

Item Score

	Correct = 1	incorrect = 0	total
References group	n11k	n12k	n1+k
Focal Group	n21k	n22k	n2 + k
Total	n+1k	n+2k	n2k

Source: Wiberg (2007)

In general the odd ratio is represented as

$$OR_{RF} = \frac{\pi R}{\pi_F} \frac{(1-\pi F)}{(1-\pi_R)}$$

Where π_F and π_R is the probability to answer an item correctly for the focal and reference group respectively. If the odd ratio is 1, it means that there is no difference between the focal group and the reference group.

Bastug (2016) stated that M.H procedures are successful in detecting basically uniform DIF but it might display or exhibit misleading result in an attempt to capture non uniform DIF particularly when using a more complex model. M.H. statistics is only suitable for classical test theory model. This is one of the major disadvantages of M.H statistics in detecting DIF. M.H has the advantage over other methods of detecting DIF. It tests for both statistical significance and effect size in order to avoid detecting items with little/greater practical significance erroneously, such as flagging DIF items. Secondly, M.H procedure is one of the most widely used methods of detecting DIF due to its simplicity and practicality.

Teodora (2013) investigated the sensitivity of Mantel Haenszel and one parameter logistic regression model in detecting differential item functioning in reading comprehension test and population of 1,925 grade six pupils was used. Six mixed sex schools and three (3) all Girl Schools were used. The two methods used indicated that, the focal group that is boys in sex base DIF, girls in mixed Sex schools in school ownership based DIF were disadvantaged in most of the items. The study revealed that M.H statistics flagged fewer DIF item that resulted to more DIF free item while the IRT logistic regression model displayed more DIF items that produced brief test instrument.

Quesen (2016) used scores from a 60-item multiple choice Mathematics assessment administered statewide to eighth graders and examined the effect of similar versus dissimilar proficiency distributions on uniform DIF detection. Results from testing the similar- and different-ability reference groups with Students with Disabilities (SWD), focal group were compared for four models: logistic regression, hierarchical generalized linear model, the Wald-1 IRT-based test, and the Mantel-Haenszel procedure. A DIF-free-then-DIF strategy, using a Wald-2 test to identify DIF-free anchor items, were used with these methods. The rate of DIF detection was examined for both similar and dissimilar distribution groups among all accommodated scores and the most common accommodation subcategories (extended time, frequent breaks, some/all items read aloud) to see if ownership of accommodation changed the rate of items flagged for DIF. The result of the study revealed that no items were detected for DIF using the similar distribution reference group, regard greater of method. No items were detected for DIF with either reference group when the IRT-

based Wald-1 test was used. With the dissimilar reference group, logistic regression had the lowest rate of items flagged for DIF (<5%), Mantel Haenszel flagged 8-15% of items, and hierarchical generalized linear model flagged 23-38% of items for DIF. Forming focal groups by accommodation ownership did not alter the pattern of DIF detection observed among models. This study found that creating reference group to be similar in ability to the focal group by purposefully sampling from the reference population might control the rate of erroneous DIF detection for SWD.

Ndifon, Umoinyang and Iduku (2010) investigated whether the 2010 Junior Secondary Certificate Examination (JSSCE) in Mathematics exhibits sex, school location and school ownership differential item functioning (DIF) in the Southern educational zone of Cross River State. Two DIF detection methods were used to identify items that exhibited DIF in 2010 JSSCE in Mathematics. The findings showed that there was no significant Sex differential item functioning as none of the detection methods identified items that function differentially between males and females. There was a significant school location differential item functioning as the Mantel-Haenszel Statistics detected two items that function differentially against urban students while the Scheuneman chi-square (SSX2) detected one item that functions differentially against urban students. Also, there was a significant school ownership differential item functioning as the Mantel-Haenszel statistics identified two items that did not favour public school students. On the other hand, the Scheuneman chi-square (SSX2) did not flag any item as functioning differentially between public and private school students. It was concluded that some items in a test used locally could exhibit significant DIF and it was recommended that DIF studies should be conducted by test developers on their test so that the items exhibiting Differential Item Functioning (DIF) could be revised or eliminated so that fairness can be enhanced.

Standardized P Difference Method of Detecting DIF

This is a non-parametric contingency table approach in detecting DIF. Standardization was propounded by Dorans and Kulick, (1986). It aims at creating a proportion difference. The approach intends to combine difference in proportion of examinees who responds to an item correctly across sub-group (reference and focal) given their levels of total test scores. Wiberg, (2007) added that there are two versions of standardization, namely; the unsigned proportion difference and the signed proportion difference. It is also regarded as standardized p-differences and root-mean weighted squared differences. The standardized approach is used to measure the effect size of DIF.

Clauser and Mazor, (1998) standardized is represented as $D_{std} \sum w\,(P_1\text{-}P_2)$

Where = proportion correct on the study item for focal group within score groups
= respective value for reference group members
W = the relative frequency of standardization group member (focal group within group)
Wiberg, (2007) represented standardization as

$$STDP - DIF = \frac{\sum_{k=1}^{k} n_2{+}k\left(\frac{n11k}{n1{+}k} - \frac{n21K}{n2{+}k}\right)}{\sum_{k=1}^{k} n_2{+}k}$$

This indicates that the DIF is either greater than 0.10 or greater than (minus) -0.10. Therefore, there is high relationship between Mantel Haenszel because the M.H used, observes test scores and matching variable. It can only detect uniform DIF like the M.H method.

One of the advantages of standardized method is that, it is easy to work with and it gives suitable result (Wiberg, 2007). It gives a good description in explaining the nature of DIF. The major demerit of standardized approach is that, it lacks an association of a test of significance (Clauser & Mazor, 1998).

Gomez-Bentio, Balluerka, Gonzalez Widaman and Padilla (2017) assessed the importance of parallel forms within classical test theory and DIF. The study focused on comparison from the total test scores to each of the item developed during test construction, analysis was based on the performance of a single group of individual on parallel items designed to ascertain (measured) the same behavioural criterion by several DIF methods. 527 examinees responded to the two parallel forms of the attention deficit-hyperactivity disorded scale. The study revealed that from the 18 items, 12 items (66.66) showed probability values associated with Mantel Haenszel chi-square statistics of greater than 01 while the standardization procedures revealed that half of DIF items favoured form A and other half form B. The study also showed that DIF of behavioural indicator can provide useful information on parallism between pairs of item to complement tradition analysis of equivalent test form based on total scores using M.H and standardization method.

Delta Plot Method of Detecting DIF (Transforming Item Difficulty)

Magis and Bruno (2011) revisited the concept of Angoff's delta method in detecting DIF, Delta Plot was first propounded by Angoff & Ford 1973 using graphical approach to compute and interpret. It was criticized by Shephard, Camilli and Williams (1985). Despite the criticism according to Osterllin and Everson (2009), Penfield and Camilli (2007), Delta plot is appropriate for mathematical testing, testing adaptation processes and intelligence testing in special population. Until 2011, there was no designed approach to select an appropriate DIF flagging criteria Magis and Facon (2011) modified the delta plot. However, Delta plot was regarded as Transform Item Difficulties (TID) method where the Delta plot compare for each item the proportion of correct responses in each group.

These correct proportion is known as P value, while the proportion for correct item j in group g is represented as p_{jg}, g is denoted as O for reference group and g denoted as 1 for focal group.

Magis and Facon (2011) opined that the proportions stated above are transformed into delta scores in two steps: firstly, P_{jg} are transformed into normalized Z scores Z_{jg}. The Z score is the deviate of the standard normal distribution with the lower tail probability $1 - P_{jg}$. It is obvious that easier items will get lower Z-scores than difficulty items (Facon & Nuchadee, 2010). Delta plot does not require advanced computer software unlike the IRT methods; the delta plot is a conceptual sample. Van Herwagen, Farran and Annaz (2011) opined that delta plots are potentially useful when sample sizes are not large, it is certain that delta plot is a method for DIF studies with small samples of respondents.

Magis and Facon (2011) explained that identifying DIF with Delta plot is important and pairing the Delta scores D_{jg} can be graphically represented on a scattered plot with reference group on horizontal axis and focal group on vertical axis which is referred to as Delta plot. Margis and Facon (2011) pointed out that when there is no DIF item, narrow eclipse along major or principle axis and the relationship between the paired Delta score is likely to be large, however, if DIF item is present, Delta plot will clearly depart from the narrow eclipse. Therefore, the DIF item will be visible easily. Flagging item as DIF or without DIF, it is necessary to compute the distance of each item from the major axis of eclipse D_j (Magis & Fucon, 2011).

Kashere Journal of Education 2022, 3(1): 146-158 **ISSN: 2756-6021 (print) 2756-6013 (online)**
DOI: https://dx.doi.org/10.4314/kje.v3i1.19 **Creative Commons Attribution License (CC BY 4.0)**
Aituariagbon, K. E. and Osarumwense, H. J.

The distance D_j between Delta plot (D_{jo}, D_{ji}) and the major axis is given as:

$$Dj = \frac{b\Delta jo + a - \Delta j1}{\sqrt{b^2+1}}$$

Where

$$b = \frac{s_1^2 - s_0^2 + \sqrt{(s_1^2 - s_0^2)^2 + 4s_{01}^2}}{2s_{01}}$$

$$a = \bar{x}_1 - b\bar{x}_0$$

Where $\bar{x}$, = Sample mean,
S^2 = Sample variance
S_{01} = Sample covariance of the Delta Score Δjg

Large perpendicular distance indicates large departures of Delta points from the major axis of the eclipse which indicates the presence of DIF. Items that is large positive distance are located above major axis are consequently easier for respondents in the reference group. On the other hand, negative perpendicular distance refers to items which are easier for respondents in the focal group. The absence of DIF shows that very small distance proposes that the item difficulties are similar across groups.

The above assumptions are known as classical delta plot, Magis and Fucon (2011) modified the delta plot with the quest to improve on the classical Delta plot. They assume that the null hypothesis, that is absence of DIF, the delta point arised from a bivariate normal distribution (Johnson & Wichern, 1998) although the bivariate normal distribution assumption obtainable and considerate, but it is difficult to assess the biveriate normality assumption with few items, it become imperative to further investigate the perpendicular distance of absence of DIF.

The modified delta plot retained the classical delta plot with a slight adjustment, the location of the eclipse is shaped through b, the significant level will increase the threshold and reduce the mistake of non DIF items. Therefore, the relationship between the paired delta scores decreases, the covariance become increasingly smaller than the variance leading to increase in the threshold. Furthermore, with lower relationship, the delta plot produces a wider eclipse (Magis & Fucon, 2011). If an item displays a large perpendicular distance in relative to other items, DIF will occur, that is, if the delta plot are all close to the major axis, the relationship between the delta score will be high and relatively small perpendicular distance will be enough to identify DIF items. Similarly, when the delta plot are scattered around the major axis, the delta scores correlation will be smaller and larger perpendicular distance will be necessary to separate DIF from non DIF. The simple implication is that there will be an increase in the threshold value (Magis & Fucon, 2011).

Muiz (2001) explained the advantages of delta plot as: "it is easy to explain, and intuitive, delta plot takes potential differences in group proficiency distribution into account in the computation of the perpendicular distance Dj. The major axis of eclipse can be close to shifted from or rotate around the identity line. These three situations correspond with the cases of equal distribution, unequal average proficiency and unequal dispersions of proficiency. The major axis automatically adjusts for difference in distributions, so that the perpendicular distances are not affected by any differences between the groups of respondents. In addition, the computation of delta value, major axis and perpendicular distance is so straight forward and does not require intensive computer implementation (Magis & Fucon, 2011).

Delta plot is not without challenges; a lot of researchers disregard the delta plot. Holland and Thyler (1988) claimed that delta plot is only built on comparison of proportion

correct value, it is designed to identify between group differences in item difficulties that is uniform DIF, if the items display group differences in item discrimination that is non-uniform DIF, then it cannot be detected by delta plot method. This makes the methods similar to the Mantel Haenszel method.

Secondly, Angoff (1993) opined that Mathematically, DIF issue occurs with the delta plot method when the proportion P_{jg} are exactly equal to 0 that is, when all responses are incorrect or 1 that is, when all responses are correct. Therefore, the delta scores become infinite. Angoff (1993) uses a range of 0.05 – 0.95 while (Magis and Fucon, 2011) used a range of 0.001 to 0.999. In addition, Facon and Nuchadee (2010) explained that all the proposed classification rules irrespective of the modified delta plot, suggest a fixed quantity which does not make reference to a pre-specified significant level rather they tend to act as measures of effect size. The value 1.5 is clearly related to thresholds for distinguishing between negligible, moderate and large DIF effect size.

Finally, according to Jodoin and Gierl (2001), with large samples or large tests, it is expected that delta points will lie very close to major axis relatively small depart from the axis as indications of DIF. However, the fixed, value threshold does not agree with this, it is seen that even items that are large cannot be detected with this method. Therefore, threshold fixed value rule is a serious source of potential conservativeness.

Abedalaziz, Leng and Alahmadi (2016) applied transformed item difficulty approach in detecting sex-related DIF using multiple choices mathematic ability test. The study showed that female examinees have a statistical significance and consistent advantage over male on items involving algebra but the male examinee reveal a greater consistent advantage on items involving geometry and measurement, number and computation, data analysis and proportional reasoning. Therefore, sex difference in mathematics test item is related to the content.

Sasaki (1991) comparing two approximate techniques of detecting DIF in language, using English as a second language placement test when the group sample size is not large enough to use other methods. However, the study applied the delta plot method using the one parameter Rasch method and Scheuneman's chi-square method. The study used 844 foreign students who took English as a second language with 61 native language background and 76 academic specializations, 262 Chinese speaking examinees. The study revealed that there were only marginal differences between DIF item detected by Delta plot and Scheuneman Chi-square procedure. The delta plot approach detected fewer DIF items with little variety than Scheuneman's method. It was also showed that Delta plot detect easier items with smaller differences in p-value between the two groups while scheuneman's approach detect items with the opposite features.

Economics is a subject in the school system. The role of Economics as a subject in the senior secondary school cannot be compromised, despite its removal as core subject and replacement with Civic Education. Yet there is no reduction in the enrolment at the Senior Secondary Certificate Examination S.S.C.E. Economics is unique because it involves some subjects in the Junior Secondary School such as Social Studies. Also, at the tertiary level, Economics is the basic for the Social Science particularly Business Administration, Banking and Finance, Financial Management and Statistics. Economics is a pre-requisite for admission into social science courses in higher institutions of learning.

Mantel Haenszel was significant (DIF) when the MH chi-square were greater than MH critical value (vice visa). When MH alpha were greater than 1, the item favour reference group, MH less than 1 the item favour focal group. STDP was significant (DIF) when the STDP-DIF value was either greater than the 0.10 or less than -0.10 (>0.10 0< 0.10). Value of STDP-DIF greater than + 0.1 indicated favour reference group

while STDP-DIF value greater than -0.1 indicated favour focal group. Transform item difficulty Dj greater than plus one or greater than minus one. It showed significant (DIF) >+1.0 or < -1.0) favour focal group. >-1.0 favour reference group.

Statement of the Problem

Researchers who work on DIF, are interested in testing programmes with high stake which involves comparison of two groups at a time (reference and focal group) such as male and female, private and public schools, urban and rural. There is however a number of methods for detecting DIF, these methods are derived from different mathematical theorems, hence, the computation procedures differ. The computation procedures are not of much concern to the psychometricians as the divergent of the results obtained. DIF computer using different methods could slow that some items flagged as functioning differentially by another method, hence the need for this study to determine the suitability of two non-parametric methods of detecting DIF. Secondly to consider if the Mantle Haenzel method will flagged more items than the standardized p-difference techniques. There are also parametric approach of detecting DIF using different methods such as logistic regression, item characteristic curve (icc), log linear model, lord chi-square methods, however, this study will only consider the difference in the non-parametric approach such as: transform item difficulty, standardized p-difference and Mantel Haenszel techniques to determine the stability of one method over the other method in flagging test items.

Research Question

To guide the study, the following research questions are posed:

1. Do SSCE 2019 Economics multiple choice items function differentially by sex, using Mantel-Haenszel technique?
2. Do SSCE 2019 Economics multiple choice items function differentially by sex, using Standardized–p difference technique?
3. Do SSCE 2019 Economics multiple choice items function differentially by sex, using Transformed Item Difficulty (Delta Plot) technique?

Methodology

The research design adopted for this study was survey design. This design was chosen because the study sought to sample opinion from respondents using questionnaire and to investigate the non-parametric methods used in detecting DIF to analyze SSCE 2019 Economics multiple choice items. The population of this study consisted of 14,400 SS2 Economics students in all the senior secondary schools in Edo South Senatorial District. There are seven Local Government Areas in Edo South Senatorial District, namely; Egor, Oredo, Ikpoba-Okha, Ovia South West, Ovia North East, Orhiomwon, Uhunmwonde Local Government Area, each with numerous public and private secondary schools.

The sample size of 1,440 senior secondary school class II Economics students was selected for this study. The stratified random sampling technique was adopted.

Firstly, use strata to select urban Local Governments Areas from the seven (7) local governments Areas in Edo South Senatorial District namely; Egor Local Government Council Area, Oredo and Ikpoba-Okha Local Government Areas.

Secondly, the simple random sampling technique was used to select six (6) public schools each from the rural Local Government Areas which are Egor, Oredo and Ikpoba-Okha Local Government Areas. Therefore, eighty (80) students were selected from each school made up of 571 males and 869 females. Thirdly, Cluster sampling technique was adopted by using students in all the arms of SSS3 intact. Any intact group of similar characteristics is a cluster (Omorogiuwa, 2006). The intact class comprisesd male and female students; students.

The instrument was validated by experts in Economics and Educational Measurement and Evaluation. The reliability of the instrument was determined using Cronbach Alpha statistics which was 0.72. The

instrument was administered to SSS3 students in public schools. The data was analyzed using Mantel Haenszel, standardized p-difference and transformed item difficulty to determine the DIF.

Results

Table 1: Mantel-Haenszel Differential Item Functioning indices by Sex for SSCE 2019 Economics Objective Examination

Number of item	Number of item flagged (DIF)	Number of item not flagged n (Non DIF)	Number of item in favour of male	Number of item in favour of female
60	45	15	16	29
	75%	25%	35.6%	64.4%

The result in the table 1 above show that using the Mantel Haenszel method of detecting DIF for sex, out of the 60 multiple choice test items in the SSCE 2019 Economics test items, 45 items (75%) functioned differentially by Sex with calculated M-HCHISQ values range from 4.14 to 109.72 which were greater than critical value of 3.84 given one degree of freedom at 0.05 alpha level, while 15 items (25%) did not function differentially. Moreover, 16 items (35.6%) favoured male students while 29 items (64.4%) favoured female students, these are gotten when MH alpha>1. then the item is in favour of reference group (male); MH alpha<1. Then, the item is in favour of focal group (female); MH alpha=1 neither reference nor focal group.

Table 2: Standardized Differential Item Functioning indices by Sex for SSCE 2019 Economics Objective Examination

Number of item	Number of item flagged (DIF)	Number of item not flagged n (Non DIF)	Number of item in favour of male Reference group	Number of item in favour of female focal group
60	14	46	3	11
	23.3%	76.7%	21.4%	78.6%

The result in the table 2 above show that using the Standardized p-difference method of detecting DIF for sex, out of 60 items, 46 items (76.7%) do not possess DIF while 14 items (23.3%) possess DIF The item is interpreted as a DIF item if the standardized p-difference value is either >0.10 or <-0.10. A value of STDP-DIF greater than +0.1 indicates DIF favoring reference group (male), whereas a value of STDP-DIF greater than -0.1 indicates DIF favouring focal group (female). Therefore, out of 14 items that functioned differentially, three items (21.4%) favoured male group while 11 items (78.6%) favoured female group.

Table 3: Transformed Item Difficulty (Delta Plot) Differential Item Functioning Indices by Sex for SSCE 2019 Economics Objective Examination

Number of item	Number of item flagged (DIF)	Number of item not flagged n (Non DIF)	Number of item in favour of male	Number of item in favour of female
60	15	45	1	14
	25%	75%	6.7%	93.3%

The result in the table 3 above shows that using the Transformed item difficulty (Delta plot) method of detecting DIF for sex, out of 60 items, 45 items (75%) do not possess DIF while 15 items (25%) possess DIF. The items with values greater than +1 or greater

than -1 revealed DIF. A value of Di greater than one unit indicates DIF favoring females, whereas a value of Di greater than minus one unit indicates DIF favouring males. Therefore, out of the 15 items that functioned differentially, one item (6.7%) favoured male while 14 items (93.3%) favoured female.

Discussion

The findings on question one revealed that 45 items function differential with 75% and 15 items did not function differentially with 25%, 16 items in favour of male with 35.6% and 29 items in favour of female with 64.4%, this finding disagree with Teodora (2013) which stated that MH statistics flagged fewer items, however, this study showed that M.H flagged more item (45/75) also with respect to sex male students are more disadvantageous using MH statistics company to female. This study also disagrees with the findings of Ndifon, Umoinyang and Iduku (2010) stating that none of the items functioned differentially with respect to sex. This study, however disagree with Quesen (2016) whose study revealed that MH only displayed 8 items and 18% but higher than 5%, although M.H was compared with four other methods parametric and non-parametric. In this study, using M.H displayed more DIF items which is in total disagreement with findings of Quesen.

Findings on Research Question 2 revealed that standardized p-difference flagged only 14 items with 23.3% and 46 items were not flagged, this method was more favourable to female with 11 items (78.6%) than male 3 items (21.4%). This study agrees with Wibeng (2007) who stated that standardized p-difference gives suitable result, considering the result above, it is comparable more suitable and consistent by displaying fewer items. This finding also agree with Gomez-Bentio, Balluerka, Gonzalez Widaman and Padilla (2017) whose ascertained that standardized p-difference displayed fewer items compared to M.H and chi-square statistics, also standardized p-difference show DIF of behavioural indicator, which agree with concept of sex as variable to determining how they respond to the item, male and female examinee.

The findings on question three 3 showed that transformed item difficulty Delta plot flagged 15 items (25%) and did not flagged 45 items (75%(with 14 items (93.3%) in favour of female and only 1 item (6.7%) favour male examinee. This finding agree with Abddalaziz, Leng and Alahmadi (2016) who state that sex difference in response to test items is related to the contest of the test. However, the male and female examinee responded differently considering more items in favour of female and less item in favour of male. This finding also agree with Sasaki (1991) whose compared Delta plot with chi-square and concluded that Delta plot flagged fewer items, considering this study, Delta plot also flagged few items compared to the M.H statistics.

Conclusion

Based on the findings of the study, the following conclusions were reached: items administered by NECO 2019 Economics Multiple Choice Questions using three non-parametric methods to determine if the item functioned differentially by sex showed that Mante Haenszel method is not a suitable method in detecting 2019 Economics Multiple Choice items prepared by NECO because, it displayed more items. Secondly, standardized p-difference and Delta plot/transform item difficulty are suitable methods in detecting differential item functioning.

Recommendation

Test item cannot be ascertained without acknowledging the process of item analysis in order to determine if the test item is either fair, bias and items functioning differentially. Hence, it is recommended based on the findings that:

1. Educational measurement expert should use other non-parametric approach in detecting DIF, this will help to determine if other non-parametric methods will be more suitable than the

standardized p-difference and transformed item difficulty delta plot.

2. DIF analysis should be used during trial testing of achievement test by National Examination Body such as West African Examination Council (WAEC), National Examination Council (NECO), National Business and Technical Examination Board (NABTEB) and Joint Admission and Matriculation Board (JAMB) to overcome consistent measurement error in testing.

References

Abdulaziz, L. (2016). Differential item functioning in on line learning instrument (EPFUN) *Creative Education*7, 180 – 188.

Acar, T. & Kelecioğlu, H. (2010). Comparison of Differential Item Functioning Determination Techniques: HGLM, LR and IRT-LR. *Educational Sciences: Theory & Practice*, 10 (2), 639-649.

Adediwura A. A. (2013). A Comparative Study of Item Response Theory and Generalized Linear Model Methods of detecting differential item functioning in dichotomous test. Research.

Angoff W.H. (1988). Validity: An evolving concept. In H. Wainer & Braun. H. (EDS) *Test validity* p 19-32 Hillsdale. N Erbaum.

Angoff, W. H. & Ford, S.F (1973). Item race international on a test of scholastic aptitude.
Journal of Educational measurement 10, 95 – 106.

Bastug, O. Y. O. (2016). A comparison of four differential items functioning procedures in the presence of occult dimensionality. *Educational Research and Reviews.* 11 (3) pp. 1251– 1261.

Clauser, B. E. & Mazor, K. M. (1998). Using statistical procedure to identifying differentially functioning test items. *Educational Measurement: Issues and practice.* 17(1), 31 – 44.

Camilli, G. & Sheppard, L. A. (1994). Methods for Identifying Biased Test Items. Thousand Oakes, CA; Sage.

Dorans, N. J. & Kullick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on scholastic aptitude Test. *Journal of Educational Measurement.* 23, 355 – 368.

Emaikwu, S. U. (2012). Issues in test item bias in public examination in Nigeria and implication for testing. *International journal of Academic Research in Progressive Education and Development.* 1(1) 175-187.

Facon, B. & Nuchade, E, M. L. (2010). An item analysis of Raven's colored progressive matrices among participants with down syndrome. *Research in Developmental Disabilities.* 31, 243-249

French, B. F., Hand, B., Nam, I. H. & Vazquez, J. V. (2014). Psychological Test and Assessment Modeling. 5(3), 275-286.

Gomez – Benito, J. BaIluerka, N., Gonzalez A., Widaman, K. & Padilla, J. (2017). Detecting differential item functioning in behavioural indictors across Paralle forms. *Psicothema.* 29(1), 91 – 95.

Gierl, M. J., Jodoin, M. G., & Ackerman, T. A. (2000). Performance of Mantel Haenszel, simultaneous items bias test of logistic regression when the proportion of DIF is large. Paper presented at the annual meeting of the American Educational Research Association (AERA) New Orleans, Lousiana, USA

Johnson, R. A. & Wichern, D. W. (1998). *Applied Multivariate statistical*

analysis. Person prentice upper saddle river new jersey (Edition).

Magis, D. & Facon, B. (2011). Angoff's delta method revisited; improving DIF detecting under small sample. *British Journal of Mathematical and Statistical Psychology*.

Mantel, N. & Haenszel, W. (1959). Statistical Aspect of the analysis of data from retrospective studies of disease. *Journal of National Cancer Institute. 22, 719-748.*

Ndifon, B. O., Umoinyang, I. E. and Iduku, F. O. (2010). *Differential Item Functioning of 2010 Junior Secondary School Certificate Mathematics Examination in Southern Educational Zone of Cross River State, Nigeria.* Paper presented at the 7th annual conference of Nigeria Association of Educational Psychologists Zaria, Nigeria.

Ogbebor U. & Onuka, A. (2013). Differential items functioning methods as item bias indication *Educational Research* 4(4) 367-373 online at http://www.intersjournals qrs/ER.

Omorogiuwa, K. O. (2006). *Research and Applied statistics for the behavioural sciences: An introduction*. Benin: Mindex press limited.

Osterlind, S. J., Everson, H. T. (2009). *Differential item function* (2nd ed) Thousand oaks, CA: sage.

Quesen, S. (2016). *Differential item functioning for accommodated students with Disabilities: Effect of differences in proficiency distributions*. PhD Dissertation, School of Education, University of Pittsburgh.

Sasaki, M. (1991). A comparison of two methods for detecting in an ESL placement test. Language testing, 8(2) 95 – 111.

Sheppard, L. A., Camilli, G & Williams, D. M. (1985). Validity of approximation techniques for detecting items bias. *Journal of Educational Measurement.* 22, 77 – 105.

Theodora, M. S. (2013). Differential item functioning detection in reading comprehension Test using mental Haenszel, item response theory and logical data analysis. *The International Journal of Social Sciences.* 1(14). www.tijoss.com

Van Herwegen, J., Farran, E. & Annaz, D. (2011). Item and error analysis of ravern's colored progressive matrices in Williams syndrome. *Research in Developmental Disabilities.* 32, 93 – 99.

Wiberg, M. (2007). *Measuring and Detecting Differential item functioning in criterion referenced test. A theoretical comparison of methods.* UMEA university press.

Yaghoobi, M. (2016). Fairness and Bias in language testing. *International Journal of Research in Linguistics, Language Teaching and Testing*. 3(1), 136 -143.

Zhang, Y. (2015). *Multiple way to detect differential item functioning in SAS.* Education Testing service (ETS)

Zumbo, B. D. (2007). Three Generation of DIF Analysis: considering where it has been where it is now and where it is going. *Language Assessment Quarterly.* 4(2), 223-233.