Assessment of Psychometric Qualities of Gender Performance in Basic Education Certificate Examination in Mathematics Multiple Choice Test Items

[1]Ukamaka. E. **AKUCHE**
akucheukamakae@gmail.com
+234 805 542 5576

&

Taiwo R. **ALIYU**
aliyutaiwo2013@gmail.com
+234 806 856 1852

[1&2]Department of Science Education,
Faculty of Education
Lead City University, Ibadan,
Oyo State, Nigeria

Abstract

The study was aimed at assessing the psychometric qualities of gender in Mathematics multiple choice test items in Basic Education Certificate Examination (BECE) in Oyo State. Four research questions and one hypothesis guided the study and was tested at 0.05 level of significance. An instrumentation research design was adopted. The population of this study consisted of all Junior Secondary class three students in Ibadan, Oyo State. The multistage sampling technique was employed to sample out 750 students from the population. The instrument used for data collection was the 2017 BECE Mathematics multiple choice test items. The reliability value of the items using KR_{20} was 0.78. DIF measure statistics of Winsteps was used to assess the difficulty indices of both the male and the female involved in the study and also tested the stated hypothesis at 0.05 level of significance. The finding revealed that there were incidences of gender variance in the test scores. Reliability with IRT of .98 showed high representativeness of the items but lack local independence of item, hence, it lacked unidimensionality.The test contained 19 bad items and 41 good ones, therefore it was concluded that the 19 items should either be deleted, reviewed, removed or restructured. It was therefore recommended that examination bodies should be mindful of the existence of 'item noise' which could *cause bias in gender performance in an examination. It was equally recommended that government should developed calibrated MAT items Bank.*

Keywords: Assessment, Calibrate, DIF, Item Bank, Rasch, IRT, 3-PL Model

Introduction

A desirable test is one that is simple and easy to use and is characterized by high quality of the information obtained which is usually reported as reliability and validity. Some tests are relatively straightforward, like some of those used in the education or physical sciences. A good test items should yield invariant scores. Invariance describes the 'scope of use' properties of a good test. For example, a ruler provides scores of height in inches. The 'height' scores are invariant: regardless of the ruler used, a person's height remains constant and the ruler can be used with anyone. A ruler's use is not restricted to particular groups of people and is not biased towards men or women. A ruler which is marked wrong will always give the same (wrong) measurements. It is very reliable, but not very valid.

In many fields, such as medical research, educational testing, and psychology, there will often be a trade-off between reliability and validity. But, different item response patterns can provide interesting information about the characteristics of testees. For example, testees whose parents are farmers may have more difficulty getting items that reflect Mathematical Arithmetic such as loss and profit, discount or amount and interest than testees whose parents are business men and women who are into goods and services. This gives us information about the different testees groups. However, the failure of invariance prohibits group comparisons since the variable's (gender, location, school type or social economic status) definition changes for the different types of testees. This is a validity issue. When test data meet the assumption of unidimensionality, the data also meet the assumption of local independent (Green & Frantom, 2002).

Specific objectivity is another desirable characteristic in a test. Specific objectivity means that a person's trait is independent of the specific set of items used to measure it. For example, it shouldn't matter which ruler is used to measure a person's height; any ruler could be used and any one used would be independent of the person's height. Additionally, a test with specific objectivity would not be affected by missing data. Hence, despite missing data, the test would still be useful and provide credible information of the testees. Test with specific objectivity can be tailored to any given testee, thus permitting individually administered and precluding administration of items that are not appropriate for a particular testee.

A statistic known as ‘fit’ provides an internal mechanism for identifying inappropriate responses to the items, allowing exclusion or re-assessment of persons whose responses make no sense, that is, do not fit, according to the understanding of the construct (Green & Frantom, 2002). The basic idea that one can capitalize on is that the statistical behavior of "bad" items is fundamentally different from that of "good" items. The items have to be administered to testees in order to obtain the needed statistics. This fact underscores a point of view that tests can be improved by maintaining and developing a pool of "good" items from which future tests will be drawn in part or in whole. This is particularly true for instructors who teach the same course more than once.

Once the instructor is satisfied that the test items meet the above criterion and that they are indeed appropriately written, what remain is to evaluate the extent to which they discriminate among testees. The degree to which this goal is attained is the basic measure of item quality for almost all multiple-choice tests. For each item the primary indicator of its power to discriminate testees is the correlation coefficient reflecting the tendency of testees selecting the correct answer to have high scores. This coefficient is reported by typical item analysis programs as the item discrimination coefficient or, equivalently, as the point-biserial correlation between item score and total score. This coefficient should be positive, indicating that students answering correctly tend to have higher scores. Similar coefficients may be provided for the wrong choices. These should be negative, which means that students selecting these choices tend to have lower scores (Hambleton, Swaminathan, & Rogers, 1991)

Two theories sustained test development in Measurement and Evaluation. These are Classical Test Theory (CTT) and Item Response Theory (IRT). According to CTT, test scores can be decomposed into the true score component and error score component. The error score component is random and can be eliminated by sampling. Test constructed under the CTT are prone to item parameter variance across subpopulation of test takers. This is a major weakness of test constructed under CTT (Aliyu & Ocheli, 2013). Item Response Theory (IRT) decomposes test scores into true score component, systematic error component and random error component. The recognition of systematic error component is a deviation from CTT. Experts in Measurement and Evaluation are shifting from CTT onto the IRT model for test development. Researches have shown that the psychometric properties of test item such as the difficulty indices are stable across subpopulation of test takers in tests that are constructed under this model (Aliyu, 2015; Wagner-Menghin & Master, 2013).

The first established test theory called the Classical Test Theory (CTT) revolves around the concepts of true score, measurement error and index of test reliability. CTT relates observable trait (the test score, X) with the unobservable trait (the person's true ability on the characteristics, T) with the following equation: $X = T + E$, where E = measurement error (Osadebe, 2010). Item Response Theory (IRT), meanwhile, relates responses to test items (observable trait) to unobservable traits through models that specify how both trait level and item properties are related to person's item response (Embretson & Reise, 2000).

According to Wagner-Menghin & Master, (2013) and Aliyu (2015), the choice of appropriate model depends on the type of test items and their scoring. Another important consideration is that, in practice, the choice of models depends on the amount of data available. The larger the number of parameter is, the more data are needed for parameter estimation, thus requiring more complex calculation and interpretation. In this case, Rasch Model has some special properties that make it attractive to users. Rasch Model involves fewest parameters; therefore, it is easier to work with (Wagner-Menghin eta al, 2013; Aliyu, 2015). Wright (1990) gives more influential explanation in favor of Rasch Model compared to a three-parameter model. These two models are opposite in philosophy and in practice. The three-parameter model will adjust to adapt whatever type of data (includes invalid responses). The Rasch model however has tight standards in controlling the data. Unlike the three-parameter model, invalid responses such as guessing on item will not be accepted. It is described as unreliable person reliability. Critics of the Rasch Model often regard the model as having strong assumptions that are difficult to meet. However, these are values that make Rasch Model more appropriate in practice than the two and the three-parameter models.

In any mathematical models, it is important to assess the fit of data to the model. If item misfit with any model is diagnosed as due to poor item quality, for example confusing distractors in a multiple-choice test, then the items may be removed from that test form and rewritten or replaced in future test forms. If, however, a large number of misfitting items occur with no apparent reason for the misfit, the construct validity of the test will need to be reconsidered for curriculum development and the test specifications may need to be rewritten. Thus, misfit provides invaluable diagnostic tools for test developers, allowing the hypotheses upon which test specifications are based to be empirically tested against data.

Assessment is an essential component of learning and teaching, as it allows the quality of both teaching and learning to be judged and improved. It often determines the priorities of education, influences practices and affects learning in general. Changes in curricula and learning objectives are ineffective if assessment practices remain the same as learning and teaching tend to be modelled against the test. To this end therefore, the researcher wants to assess the psychometric qualities of the Basic Education Certificate Examination (BECE) in Mathematics multiple choice test items in Oyo State.

There are several methods of assessment for assessing fit for curriculum development purposes, such as a chi-square statistic, or a standardized version of it. Two and three-parameter IRT models adjust item discrimination, ensuring improved data-model fit, so fit statistics lack the confirmatory diagnostic value found in one-parameter models, where the idealized model is specified in advance (Frantom, Green & Lam, 2002).

Statement of the Problem

Most developed tests in Nigeria which are used for research, classroom or public examination purposes are based on Classical Test Theory (CTT). As a result of this, they are faced with some challenges like, poor precision, sample dependency and undue focus on aggregate scores that deny test developers the opportunity of determining how the testees performed on a test item. This problem may be addressed or overcome with the application of item response theory (IRT) of the Rasch model.

From the above stated problems, the researcher therefore wants to assess the psychometric qualities of gender performance in BECE in Mathematics multiple choice test items using the Rasch model of the Item Response Theory. Hence, the statement of problem if put in a question form is: What are the psychometric qualities of gender performance in BECE Mathematics multiple choice test items?

Purpose of the Study

The main purpose of the study was to assess the psychometric qualities of the BECE Mathematics multiply choice test items in Oyo state. As such, this study was set to achieve the following specific objective:

i. to compare the difficulty index of the BECE Mathematics multiple choice test items for male and female testees using IRT model.
ii. to determine the reliability of the BECE Mathematics multiple choice test items using IRT model.
iii. to determine the difficulty index of each item of the BECE Mathematics multiple choice test using IRT model.
iv. to establish whether the BECE Mathematics multiple choice test items are unidimensional.

Research Questions

This study therefore attempted to answer the following research questions.

(1) What are difficulty indices of male and female testees in BECE Mathematics multiple choice test item using IRT model?
(2) What is the reliability of BECE Mathematics multiple choice test item using IRT model?
(3) What is the difficulty index of each item in the BECE Mathematics multiple choice test using IRT model?
(4) Are the BECE Mathematics multiple choice test items unidimensional?

Methodology

This study adopted instrumentation Research Design. The population of this study comprised the entire Junior Secondary School three (JSS3) students in all 33 local government areas in Oyo State. This population was chosen because they must have covered the syllabuses. The sample size of this study was seven hundred and fifty (750) students of the population who were JSS 3 students from schools in Ibadan South West. A multi-stage sampling technique was adopted for the study. Simple random sampling technique was used to select the one LGA out of the 33 LGA in Oyo State. Purposive random sampling technique was used in selecting 10 Junior Secondary schools that have participated in BECE for at least five (5) consecutive years in the selected local government. Disproportionate/non-proportionate Stratified random sampling technique was used to select 75 students each from the ten (10) schools making a total of 750 respondents used in the study. The instrument of the research was BECEMAT-2017; Questions answered and treated in the year 2017 of BECE. The content and face validity of the instrument was established using the test blue print with the Mathematics syllabus from the ministry of Education and two Mathematics teachers who are qualified through TRCN certification respectively.

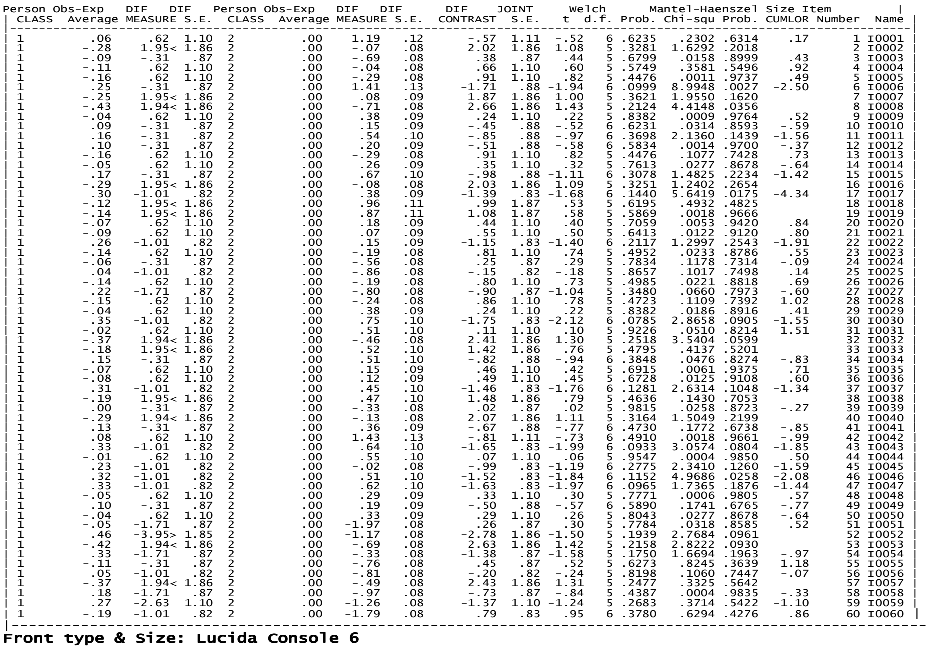
Reliability of the instrument was established with the use of Kuder-Richardson formular 20 (KR_{20}). The calculated coefficient of the reliability was 0.88 which indicated that the test items could be administered to the targeted audience.

Results

The results obtained in this study are presented and discussed. Winsteps 3.75.0 was used to answer the research questions

Research Question 1: What are difficulty indices of male and female testees in BECE Mathematics multiple choice test item using IRT model?

Table 1: DIF class specification: Male=1 Female=2 showing difficulty indices of 60 items for 750 Testees



```
Person Obs-Exp   DIF    DIF   Person Obs-Exp   DIF   DIF       DIF    JOINT    Welch        Mantel-Haenszel  Size Item        |
| CLASS Average MEASURE S.E.  CLASS Average MEASURE S.E.    CONTRAST  S.E.    t d.f. Prob.  Chi-squ Prob.  CUMLOR Number  Name |
|------------------------------------------------------------------------------------------------------------------------------|
| 1       .06    .62  1.10  2       .00   1.19  .12      -.57  1.11  -.52 6 .6235  .2302 .6314    .17   1 I0001 |
| 1      -.28   1.95< 1.86  2       .00   -.07  .08      2.02  1.86  1.08 5 .3281 1.6292 .2018           2 I0002 |
| 1      -.09   -.31   .87  2       .00   -.69  .08       .38   .87   .44 5 .6799  .0158 .8999    .43   3 I0003 |
| 1      -.11    .62  1.10  2       .00   -.04  .08       .66  1.10   .60 5 .5749  .3581 .5496    .92   4 I0004 |
| 1      -.16    .62  1.10  2       .00   -.29  .08       .91  1.10   .82 5 .4476  .0011 .9737    .49   5 I0005 |
| 1       .25   -.31   .87  2       .00   1.41  .13     -1.71   .88 -1.94 6 .0999 8.9948 .0027  -2.50   6 I0006 |
| 1      -.25   1.95< 1.86  2       .00    .08  .09      1.87  1.86  1.00 5 .3621 1.9550 .1620           7 I0007 |
| 1      -.43   1.94< 1.86  2       .00   -.71  .08      2.66  1.86  1.43 5 .2124 4.4148 .0356           8 I0008 |
| 1      -.04    .62  1.10  2       .00    .38  .09       .24  1.10   .22 5 .8382  .0009 .9764    .52   9 I0009 |
| 1       .09   -.31   .87  2       .00    .15  .09      -.45   .88  -.52 6 .6231  .0314 .8593   -.59  10 I0010 |
| 1       .16   -.31   .87  2       .00    .54  .10      -.85   .88  -.97 6 .3698 2.1360 .1439  -1.56  11 I0011 |
| 1       .10   -.31   .87  2       .00    .20  .09      -.51   .88  -.58 6 .5834  .0014 .9700   -.37  12 I0012 |
| 1      -.16    .62  1.10  2       .00   -.29  .08       .91  1.10   .82 5 .4476  .1077 .7428    .73  13 I0013 |
| 1      -.05    .62  1.10  2       .00    .26  .09       .35  1.10   .32 5 .7613  .0277 .8678   -.64  14 I0014 |
| 1       .17   -.31   .87  2       .00    .67  .10      -.98   .88 -1.11 6 .3078 1.4825 .2234  -1.42  15 I0015 |
| 1      -.29   1.95< 1.86  2       .00   -.08  .08      2.03  1.86  1.09 5 .3251 1.2402 .2654          16 I0016 |
| 1       .30  -1.01   .82  2       .00    .38  .09     -1.39   .83 -1.68 6 .1440 5.6419 .0175  -4.34  17 I0017 |
| 1      -.12   1.95< 1.86  2       .00    .96  .11       .99  1.87   .53 5 .6195  .4932 .4825          18 I0018 |
| 1      -.14   1.95< 1.86  2       .00    .87  .11      1.08  1.87   .58 5 .5869  .0018 .9666          19 I0019 |
| 1      -.07    .62  1.10  2       .00    .18  .09       .44  1.10   .40 5 .7059  .0053 .9420    .84  20 I0020 |
| 1      -.09    .62  1.10  2       .00    .07  .09       .55  1.10   .50 5 .6413  .0122 .9120    .80  21 I0021 |
| 1       .26  -1.01   .82  2       .00    .15  .09     -1.15   .83 -1.40 6 .2117 1.2997 .2543  -1.91  22 I0022 |
| 1      -.14    .62  1.10  2       .00   -.19  .08       .81  1.10   .74 5 .4952  .0233 .8786    .55  23 I0023 |
| 1      -.06   -.31   .87  2       .00   -.56  .08       .25   .87   .29 5 .7834  .1178 .7314   -.09  24 I0024 |
| 1       .04  -1.01   .82  2       .00   -.86  .08      -.15   .82  -.18 5 .8657  .1017 .7498    .14  25 I0025 |
| 1      -.14    .62  1.10  2       .00   -.19  .08       .80  1.10   .73 5 .4985  .0221 .8818    .69  26 I0026 |
| 1       .22  -1.71   .87  2       .00   -.80  .08      -.90   .87 -1.04 5 .3480  .0660 .7973   -.60  27 I0027 |
| 1      -.15    .62  1.10  2       .00   -.24  .08       .86  1.10   .78 5 .4723  .1109 .7392   1.02  28 I0028 |
| 1      -.04    .62  1.10  2       .00    .38  .09       .24  1.10   .22 5 .8382  .0186 .8916    .41  29 I0029 |
| 1       .35  -1.01   .82  2       .00    .75  .10     -1.75   .83 -2.12 6 .0785 2.8658 .0905  -1.55  30 I0030 |
| 1      -.02    .62  1.10  2       .00    .51  .10       .11  1.10   .10 5 .9226  .0510 .8214   1.51  31 I0031 |
| 1      -.37   1.94< 1.86  2       .00   -.46  .08      2.41  1.86  1.30 5 .2518 3.5404 .0599          32 I0032 |
| 1      -.18   1.95< 1.86  2       .00    .52  .10      1.42  1.86   .76 5 .4795  .4137 .5201          33 I0033 |
| 1       .15   -.31   .87  2       .00    .51  .10      -.82   .88  -.94 6 .3848  .0476 .8274   -.83  34 I0034 |
| 1      -.07    .62  1.10  2       .00    .15  .09       .46  1.10   .42 5 .6915  .0061 .9375    .71  35 I0035 |
| 1      -.08    .62  1.10  2       .00    .12  .09       .49  1.10   .45 5 .6728  .0125 .9108    .60  36 I0036 |
| 1       .31  -1.01   .82  2       .00    .45  .10     -1.46   .83 -1.76 6 .1281 2.6314 .1048  -1.34  37 I0037 |
| 1      -.19   1.95< 1.86  2       .00    .47  .10      1.48  1.86   .79 5 .4636  .1430 .7053          38 I0038 |
| 1       .00   -.31   .87  2       .00   -.33  .08       .02   .87   .02 5 .9815  .0258 .8723   -.27  39 I0039 |
| 1      -.29   1.94< 1.86  2       .00   -.13  .08      2.07  1.86  1.11 5 .3164 1.5049 .2199          40 I0040 |
| 1       .13   -.31   .87  2       .00    .36  .09      -.67   .88  -.77 6 .4730  .1772 .6738   -.85  41 I0041 |
| 1       .08    .62  1.10  2       .00   1.43  .13      -.81  1.11  -.73 6 .4910  .0018 .9661   -.99  42 I0042 |
| 1       .33  -1.01   .82  2       .00    .64  .10     -1.65   .83 -1.99 6 .0933 3.0574 .0804  -1.85  43 I0043 |
| 1      -.01    .62  1.10  2       .00    .55  .10       .07  1.10   .06 5 .9547  .0004 .9850    .50  44 I0044 |
| 1       .23  -1.01   .82  2       .00   -.02  .08      -.99   .83 -1.19 6 .2775 2.3410 .1260  -1.59  45 I0045 |
| 1       .32  -1.01   .82  2       .00    .51  .10     -1.52   .83 -1.84 6 .1152 4.9686 .0258  -2.08  46 I0046 |
| 1       .33  -1.01   .82  2       .00    .62  .10     -1.63   .83 -1.97 6 .0965 1.7365 .1876  -1.44  47 I0047 |
| 1      -.05    .62  1.10  2       .00    .29  .09       .33  1.10   .30 5 .7771  .0006 .9805    .57  48 I0048 |
| 1       .10   -.31   .87  2       .00    .19  .09      -.50   .88  -.57 6 .5890  .1741 .6765   -.77  49 I0049 |
| 1      -.04    .62  1.10  2       .00    .33  .09       .29  1.10   .26 5 .8043  .0277 .8678   -.64  50 I0050 |
| 1      -.05  -1.71   .87  2       .00  -1.97  .08       .26   .87   .30 5 .7784  .0318 .8585    .52  51 I0051 |
| 1       .46  -3.95> 1.85  2       .00  -1.17  .08     -2.78  1.86 -1.50 5 .1939 2.7684 .0961          52 I0052 |
| 1      -.42   1.94< 1.86  2       .00   -.69  .08      2.63  1.86  1.42 5 .2158 2.8222 .0930          53 I0053 |
| 1       .33  -1.71   .87  2       .00   -.33  .08     -1.38   .87 -1.58 5 .1750 1.6694 .1963   -.97  54 I0054 |
| 1      -.11   -.31   .87  2       .00   -.76  .08       .45   .87   .52 5 .6273  .8245 .3639   1.18  55 I0055 |
| 1       .05  -1.01   .82  2       .00   -.81  .08      -.20   .82  -.24 5 .8198  .1060 .7447   -.07  56 I0056 |
| 1      -.37   1.94< 1.86  2       .00   -.49  .08      2.43  1.86  1.31 5 .2477  .3325 .5642          57 I0057 |
| 1       .18  -1.71   .87  2       .00   -.97  .08      -.73   .87  -.84 5 .4387  .0004 .9835   -.33  58 I0058 |
| 1       .27  -2.63  1.10  2       .00  -1.26  .08     -1.37  1.10 -1.24 5 .2683  .3714 .5422  -1.10  59 I0059 |
| 1      -.19  -1.01   .82  2       .00  -1.79  .08       .79   .83   .95 6 .3780  .6294 .4276    .86  60 I0060 |
|------------------------------------------------------------------------------------------------------------------------------
```

Front type & Size: Lucida Console 6

Table 1 shows the DIF statistics of the Rasch model method for each of the 60 items for gender. There is incidence of variation in the gender performance. 1 represents male while 2 female. Item 1 seems difficulty for female than male counterparts. This is seen on DIF measure table with male measure value of .62logit while that of female is 1.19logit. The value of

1.19logit is higher than .62logit which means that item 1 is difficult for female but simple for male counterparts. This measure goes for the rest items and the subgroup in the table.

Research Question 2: What is the difficulty index of each item in the BECE Mathematics multiple choice test Items?

Table 2: Item STATISTICS: MEASURE ORDER

```
----------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL              MODEL|   INFIT  |  OUTFIT  |PT-MEASURE |EXACT MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE      S.E.|MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| Item |
|---------------------------------------+----------+----------+-----------+-----------+------|
|    42     64    750     1.43       .13|1.03    .4|1.28   2.1|  .03   .16| 91.6  91.5| I0042|
|     6     67    750     1.38       .13| .96   -.3| .94   -.5|  .23   .16| 91.2  91.1| I0006|
|     1     79    750     1.19       .12|1.03    .3| .92   -.8|  .17   .17| 89.3  89.5| I0001|
|    18     96    750      .96       .11| .85  -1.9| .74  -3.0|  .49   .18| 87.3  87.2| I0018|
|    19    104    750      .87       .11| .95   -.6| .92   -.9|  .28   .19| 86.3  86.2| I0019|
|    30    117    750      .72       .10|1.10   1.5|1.14   1.7| -.01   .19| 84.3  84.4| I0030|
|    15    122    750      .67       .10| .91  -1.3| .85  -2.0|  .38   .20| 83.9  83.8| I0015|
|    43    127    750      .62       .10|1.02    .3|1.09   1.2|  .14   .20| 83.2  83.1| I0043|
|    47    129    750      .60       .10|1.03    .5| .97   -.4|  .18   .20| 82.7  82.9| I0047|
|    44    134    750      .55       .10|1.01    .2|1.05    .8|  .16   .20| 82.3  82.2| I0044|
|    11    135    750      .54       .10| .90  -1.7| .87  -1.9|  .39   .20| 82.1  82.1| I0011|
|    33    137    750      .52       .10| .96   -.6| .96   -.6|  .27   .20| 82.0  81.8| I0033|
|    34    138    750      .51       .10| .99   -.2| .99   -.1|  .22   .20| 81.6  81.7| I0034|
|    46    138    750      .51       .10| .93  -1.2| .93  -1.1|  .33   .20| 81.9  81.7| I0046|
|    31    139    750      .51       .10| .88  -2.1| .81  -3.0|  .45   .20| 81.7  81.5| I0031|
|    38    143    750      .47       .10| .98   -.3| .94   -.8|  .26   .21| 81.2  81.0| I0038|
|    37    145    750      .45       .09|1.09   1.5|1.08   1.3|  .04   .21| 79.2  80.7| I0037|
|     9    153    750      .38       .09|1.03    .6|1.09   1.4|  .12   .21| 82.3  79.7| I0009|
|    17    153    750      .38       .09| .88  -2.3| .85  -2.5|  .44   .21| 82.5  79.7| I0017|
|    29    153    750      .38       .09|1.02    .3|1.00    .0|  .18   .21| 79.6  79.7| I0029|
|    41    155    750      .36       .09| .97   -.5| .99   -.2|  .25   .21| 82.0  79.5| I0041|
|    50    159    750      .33       .09| .93  -1.3| .88  -2.0|  .36   .21| 79.1  79.0| I0050|
|    48    164    750      .29       .09| .96   -.7| .94  -1.1|  .29   .21| 81.1  78.4| I0048|
|    14    167    750      .26       .09| .91  -1.9| .86  -2.6|  .41   .21| 78.0  78.0| I0014|
|    12    175    750      .20       .09| .91  -1.8| .91  -1.7|  .37   .21| 79.6  77.1| I0012|
|    49    176    750      .19       .09|1.01    .2|1.00    .1|  .20   .22| 76.8  76.9| I0049|
|    20    178    750      .18       .09| .96  -1.0| .91  -1.8|  .32   .22| 76.5  76.7| I0020|
|    35    181    750      .15       .09|1.01    .2|1.02    .4|  .19   .22| 79.3  76.3| I0035|
|    10    182    750      .15       .09|1.05   1.1|1.07   1.3|  .11   .22| 73.1  76.2| I0010|
|    22    182    750      .15       .09| .98   -.4| .94  -1.1|  .27   .22| 75.2  76.2| I0022|
|    36    185    750      .12       .09|1.05   1.2|1.02    .4|  .13   .22| 72.1  75.8| I0036|
|     7    191    750      .08       .09|1.00    .0|1.00    .0|  .22   .22| 76.4  75.1| I0007|
|    21    192    750      .07       .09|1.02    .5|1.00    .1|  .19   .22| 73.9  75.0| I0021|
|    45    205    750     -.02       .08|1.04   1.0|1.11   2.3|  .12   .22| 73.2  73.4| I0045|
|     4    208    750     -.04       .08| .93  -1.9| .91  -2.2|  .36   .22| 74.7  73.0| I0004|
|     2    212    750     -.07       .08|1.01    .2|1.01    .2|  .20   .22| 73.9  72.6| I0002|
|    16    214    750     -.08       .08| .93  -1.9| .91  -2.2|  .36   .22| 73.9  72.3| I0016|
|    40    220    750     -.13       .08|1.12   3.3|1.14   3.3| -.02   .22| 67.5  71.6| I0040|
|    26    229    750     -.19       .08| .98   -.5| .96  -1.1|  .27   .22| 69.1  70.5| I0026|
|    23    230    750     -.19       .08|1.02    .5|1.00    .0|  .20   .22| 68.4  70.4| I0023|
|    28    237    750     -.24       .08| .96  -1.4| .93  -1.8|  .32   .23| 71.2  69.6| I0028|
|     5    245    750     -.29       .08| .94  -1.9| .93  -2.1|  .34   .23| 69.6  68.7| I0005|
|    13    245    750     -.29       .08| .96  -1.5| .95  -1.5|  .32   .23| 68.3  68.7| I0013|
|    39    251    750     -.33       .08| .99   -.4| .98   -.6|  .25   .23| 68.5  68.0| I0039|
|    54    251    750     -.33       .08|1.07   2.4|1.06   1.8|  .09   .23| 64.3  68.0| I0054|
|    32    273    750     -.46       .08| .97  -1.2| .94  -2.0|  .30   .23| 64.5  65.7| I0032|
|    57    277    750     -.49       .08|1.11   4.3|1.13   4.5|  .00   .23| 59.7  65.3| I0057|
|    24    289    750     -.56       .08| .99   -.2| .98   -.6|  .24   .23| 63.5  64.1| I0024|
|     3    311    750     -.69       .08| .94  -2.8| .93  -3.0|  .34   .23| 64.5  62.0| I0003|
|    53    311    750     -.69       .08|1.15   7.0|1.19   7.5| -.08   .23| 53.9  62.0| I0053|
|     8    315    750     -.71       .08| .98   -.8| .99   -.6|  .26   .23| 63.2  61.7| I0008|
|    55    323    750     -.76       .08|1.16   8.3|1.22   9.4| -.12   .23| 50.5  61.0| I0055|
|    27    331    750     -.80       .08| .92  -4.5| .91  -4.1|  .39   .23| 66.5  60.4| I0027|
|    56    332    750     -.81       .08|1.05   2.8|1.06   2.8|  .13   .23| 55.5  60.3| I0056|
|    25    341    750     -.86       .08|1.00    .0|1.01    .3|  .23   .23| 61.7  59.8| I0025|
|    58    361    750     -.97       .08|1.08   4.9|1.11   5.5|  .05   .23| 54.8  58.9| I0058|
|    52    396    750    -1.17       .08|1.05   3.1|1.07   3.6|  .11   .22| 60.3  58.8| I0052|
|    59    412    750    -1.26       .08|1.11   6.6|1.19   8.5| -.03   .22| 50.9  59.3| I0059|
|    60    502    750    -1.79       .08|1.02    .7|1.03    .9|  .16   .21| 66.4  67.5| I0060|
|    51    529    750    -1.97       .08|1.01    .3| .98   -.5|  .19   .20| 67.5  70.8| I0051|
|---------------------------------------+----------+----------+-----------+-----------+------|
| MEAN   213.5  750.0      .00       .09|1.00    .3| .99    .2|           | 73.8  74.3|      |
| S.D.    98.1     .0      .69       .01| .07   2.3| .10   2.7|           | 10.0   8.7|      |
----------------------------------------------------------------------------------------------
```

Front type & Size: Lucida Console 7.5

In answering the research question 2 Winsteps software programme was used to calibrate the responses of the 750 testees to the 60 PAT items. The table 2 above shows the difficulty indices in the fourth column, item 42 is the most difficult item in the test. The difficulty of this item is estimated to be 1.43logits with the standard error of 0.13 while item 51 is the easiest with -1.97logits and standard error of 0.08.

Research Question 3: What is the reliability of BECE Mathematics multiple choice test item using IRT model?

Table 3–Reliability table of 60 MAT items in logit

```
          TOTAL                         MODEL         INFIT        OUTFIT
          SCORE     COUNT    MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD
-------------------------------------------------------------------------------
 MEAN     213.5     750.0        .00     .09      1.00     .3    .99     .2
 S.D.      98.1        .0        .69     .01       .07    2.3    .10    2.7
 MAX.     529.0     750.0       1.43     .13      1.16    8.3   1.28    9.4
 MIN.      64.0     750.0      -1.97     .08       .85   -4.5    .74   -4.1
-------------------------------------------------------------------------------
 REAL RMSE     .09 TRUE SD      .68  SEPARATION  7.50  Item   RELIABILITY  .98
 MODEL RMSE    .09 TRUE SD      .68  SEPARATION  7.58  Item   RELIABILITY  .98
S.E. OF Item MEAN = .09
```

Table 3 shows the summary statistics of 60 measured item. This investigated the representativeness of the items by checking the value given for item strata, item separation and item reliability. The item strata is 9.4, item separation is 7.58 while item reliability is .98. The item reliability seems good for the test. Therefore, the reliability of the BECE MAT items using IRT model is .98

Research Question 4: Are the BECE Mathematics multiple choice test items unidimensional?

Table 4: Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)

```
                                             -- Empirical --    Modeled
Total raw variance in observations    =   68.7 100.0%           100.0%
Raw variance explained by measures    =    8.7  12.6%          12.8%
Raw variance explained by persons     =    1.2   1.7%             1.7%
Raw Variance explained by items       =    7.5  10.9%          11.1%
Raw unexplained variance (total)      =   60.0  87.4% 100.0% 87.2%
Unexplained variance in 1st contrast =    3.5   5.2%           5.9%
Unexplained variance in 2th contrast =    3.3  10.0%           2.1%
Unexplained variance in 3th contrast =    3.1   9.6%           1.9%
Front type & Size: Lucida Console 8
```

The table 4 is interpreted by comparing the empirical values of the entries with the modeled value. The raw variance explained by the measures of 12.6% did not agree with the model value of 12.8%, the raw variance explained by items of 10.9% did not agree with the model of 11.1%, raw unexplained variance (total) of 87.4% did not agree with the model of 87.2%.Variance in the first item should be at least 4 times greater than the variance in the first contrast in the first contrast for unidimensionality but for this test items, it is 1.9%. Also, eigenvalue of the unexplained variance in first, second and third contrasts were 3.5, 3.3, and 3.1 respectively which were not supposed to be more than 2.0 for a unidimensional test. These values showed that BECE MAT is multidimensional which queries the construct validity of the test items. Therefore, the BECE mathematics multiple choice test items are not unidimensional.

Discussion of findings

It was observed that 25 items seemed easy for the male while 35 of the BECE mathematics test items seem difficult for the female counterparts. The most difficulty item is item 42 with 1.43logit while the easiest item is item 51 with measure value of -1.97logit. It was equally observed that 19 items of the BECE Mathematics multiple choice items are bad and should be rewritten, reviewed or removed. The items are 42, 18, 31, 17, 14, 45, 4, 16, 40, 5, 57, 3, 53, 55, 27, 56, 58, 52, and 59. These items are measuring other things other the construct, so they construct irrelevant. They were selected using the bench mark of infit and outfit of MNSQ and ZSTD of .6 -1.2 and -2 +2 respectively.

Conclusion

The study concluded that the BECE Mathematics multiple Choice test items of 2017 are not unidimensional, they are construct irrelevant. In other words, the items have multidimensional traits which could have great influence on the performance of the students.

Recommendations

It is therefore recommended that the:

i. nineteen (19) bad items identified should be reviewed or removed and

ii. method of IRT model should be adopted in test development of all kinds of items in any subject.

References

Aliyu, R.T., & Ocheli, E.U. (2013). Development and validation of college Mathematics Test using Item Response Theory (IRT) models. *A Delta Journal of Educational Research and Development 12(1), 130-140.*

Aliyu, R.T. (2015). Development and Validation of Mathematics Achievement Test using the Rasch Model. *An Unpublished Ph.D Thesis of the Delta State University, Abraka.*

Aliyu, R.T. (2015). Construct Validity of Mathematics Test Items using the Rasch Model. *An International Journal of Social Science and Humanities Research. 3(2), 22-28*

American Psychological Association (APA) (2012). Concise Rules of APAStyle. *Washington, DC:APA publications.*

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? In Smith E.V. & Smith, R.M. (Eds.), *Introduction to Rasch measurement* 143-*166. Maple Grove, MN: JAM Press*.

Converse, J.M., & Presser, S. (1986). Survey questions. Newbury Park, CA: DeVellis,

R.F. (1991). Scale development: *Theory and applications. Newbury Park, CA: Sage.*

Embretson. S.E. & Reise. S.P. (2000). Item response theory for Psychologists. *Mahwah, New Jersey: Lawrence Erlbaum Associates.*

Frantom, C.G., Green, K.E., & Lam, T.C.M. (2002). Item grouping Effects on invariance of attitude items. *Journal of Applied Measurement, 3*, 38-49.

Golino, H. F, Gomes, C. M. A., Commons, M. L., & Miller, P. M. (2012). The Construction and Validation of a Developmental Test for Stage Identification: Two Exploratory Studies. *A Journal of the laboratory for cognitive Architecture Mapping (Laico) Universidade Federal de Minas Gerais Brazil, 1- 43.*

Green, K.E. & Frantom, C.G. (2002). Survey Development and Validation with the Rasch model. *A paper presented at the international conference on questionnaire, development, evaluation and testing, Charleston, SC, November 14- 17, pp 3-8*

Hambleton, R.K., Swaminathan, H, & Rogers, H.J. (1991). *Fundamentals of Item Response Theory. Newbury Park, CA: Sage Press*

IRT from SSI: (2003). BILOG-MG, MULTILOG, PARSCALE, TESTFACT edited by Math; Ida du Toit SSI. *Scienfic software international.*

Odili, J.N., Osadebe, P.U. & Aliyu, R.T. (2015). Assessment of Stability of Item Parameter in a Mathematics Achievement Test Under The Rasch Model. *Journal of Association of Educational Researcher and Evaluators of Nigeria (ASSEREN), 1(1), 1-8*

Osadebe, P.U. (2010). Construction and Validation of Test Items. *An unpublished lecture note, Delta state university.*

Wagner-Menghin, M.M, & Master, G.N (2013). Adaptive Testing for Psychological Assessment: How many items are enough to run an Adaptive Testing Alogarithm?. *Journal of Applied Measurement 14(2), 106-117*