CONSTRUCT VALIDITY AND RELIABILITY OF BASIC EDUCATION CERTIFICATE EXAMINATIONS IN NATIONAL VALUES EDUCATION IN BENUE STATE – NIGERIA

NASELA PILA
Department of Educatoona Foundations, Faculty of Education,
Benue State University, Makurdi
and
J. G. ADEWALE
Institute of Education, University of Ibadan, Ibadan

Abstract
This study assessed construct validity and reliability of Basic Education Certificate Examinations (BECE) in national values education (NVE) for 2017. Test bias with respect to testees' gender and school location was used to detect construct validity. Ex-post facto research design was adopted while Item Response Theory was employed for reliability estimation. Two research questions were answered. The study population was Junior Secondary School (JSS3) students in Benue State. Eight local government areas from two senatorial districts (4 from each) were randomly sampled; 1500 JSS3 students from 48 secondary schools across rural and urban settlements were selected using stratify sampling. National-BECE NVE-2017 was the instrument used to collect data which were analysed using exploratory structural equation modelling (ESEM) and IRT-Logistics Parameter. The findings revealed that: National-BECE NVE-2017, with 2-underlying factors, had no construct validity. It was biased to testees' subgroups (gender and school location). The two-factor model describing examinees' performance (construct validity) in National-BECE based on gender and school location differences was viable. Two-underlying factors have marginal reliability coefficient: (F1) r=0.8550695,(F2) r=0.6022473 implying that only F1 is highly reliable and indicating high internal consistency of items. It was recommended that public examining bodies should establish test construct validity and reliability before administration.
Keywords: construct validity, reliability, national values education,

Introduction

The quality of a measurement instrument largely depends on the degree to which it produces consistent accurate results to inform decisions on the measurement outcomes. As a public examination at basic level of education, National Basic Education Certificate Examination (BECE) ought to be reliable and valid for effective promotion and placement of students into senior secondary schools. The focus of this study is to detect test biases based on testees' subgroups (gender and school location) in BECE 2017 National Values Education (NVE) conducted by National Examinations Councils (NECO) in Benue State using exploratory structural equation modelling (ESEM). The study will also estimate the test's reliability using item response theory (IRT).

Tests are psychological measurement instruments which determine the level at which learners have acquired desired behaviours. Results from tests on learners' behaviour inform decisions on learning outcomes for admission, certification, promotion and placement (Nworgu, 2011). Decisions based on test scores are considered objective only when the tests are valid, in content and construct, and are reliable. That is why it is necessary to conduct psychometric analysis of a test before administration. Item response theory has been used over the years to achieve test quality. It is a set of models which relates the probability of reaction by a testee with a given ability to the characteristics of the item constructed to gather data and ascertain the extent to which an individual acquired a change in behaviour. IRT seeks to model the relationship between a testee's traits and the probability of the testee answering a particular item correctly (XinmingAn & Yung 2014). According to Nenty (2015), IRT attempts to estimate the test parameters, explain the process and predict the outcome of a given measurement for validity purposes.

This study focuses on the construct validity of BECE as a public examination conducted at basic education level. Construct validity is the degree of accuracy of inferences made on the basis of measurement, mostly using test scores. It is all about whether a test reflects a sample of behaviour(s) under measure. It concerns the extent to which empirical evidence and theoretical measure support the sufficiency and appropriateness of inferences based on the observed test scores (Weiland, Durach, Kembro, & Treiblmairer, 2017). A test with acceptable construct validity cannot produce scores for inferences that favoured testees on the basis of their subgroup differences even when it has multiple dimensions or factors.

Chima (2014) observed that most public examining bodies in Nigeria produced scores on individuals' academic achievement that do not reflect the testees' true ability. Chima doubted the validity of assessment scores generated on individual students for promotion, placement and certification by WAEC, NECO and other public examining bodies in Nigeria. Chima's assertion creates doubt whether BECE, as one of the public examination bodies in Nigeria, is using invalid and unreliable instruments to generate scores on learners' achievement.

Sideridis, Tsaousis, and Al-Sadaawi (2018) assessed the construct validity of mathematics achievement tests in Saudi Arabia and established the degree of test bias at the university level. Data for the study was collected on 2881 students who sat for national mathematics examinations as entry requirements for admission into public universities in the country. The multilevel structural equation modelling (MSEM) was used for data analysis. Using the one parameter logistics (1PL) model, four factors were discovered to underlie the test. These were: numbers and operations, algebra and analysis, geometry and measurement, statistics and probabilities. Achievement in the underlying factors favoured some university levels more than others based on year of establishment.

Bandele and Adewale (2013) investigated the comparative analysis of reliability and validity coefficients of WAEC, NECO and NABTEB mathematics tests 2012. They

found that NECO had low reliability coefficient (r = .512), WAEC (r = .815) just like NABTEB (r = .752) has higher reliability coefficient at 0.05 level of significant. Their study adopted CTT approach for reliability estimation while the present study preferred IRT approach. The survey studied a sample of 720 final year technical college students in Nigeria. This implies that NECO questions have been reliable over a period of time.

Test reliability is the degree to which a test consistently measures certain abilities or traits. It gives the confidence that the test will yield the same result if it measures the same trait across the sampled population after repeated administrations (Eluwa, et al. cited in Onu, 2014). It is the extent to which an instrument consistently measures what it purports to measure without producing different results under the same condition. Culligan (2005) described reliability as a measure of the consistency of an instrument on a particular population sample at a given point. Under item response theory (IRT), binary response items command wide applications among researchers in measurement and evaluation (De Ayala, 2009).

Under IRT model, reliability is conceptualized as '*test information*' conditioned by the ability level being measured. It is possible that some items in a test measure certain abilities adequately while others may not. Therefore, item information function (IIF) examines item usefulness for assessing testees' ability levels. Test items are said to be very informative at a point the slope of item characteristic curve (ICC) is steepest. This happens when the difficulty of an item is relatively high and discrimination is close to testee's ability (θ). It is expected that a valid and reliable test item should possess difficulty and discrimination as item parameters. In line with this assumption, testees with higher ability levels are expected to answer difficult items correctly. This differentiates them from those with low ability who may not rightly respond to such test items. Idang (2009) in Ogbebor (2012) stated that reliability is relative to the standard deviation of a test and to the p-value of the test items, which are dependent upon the traits of particular examinees and the characteristics of the test.

One of the basic interpretations for test validity is that, a test consistently measures the psychological ability. Reliability is a measure of consistency of the application of an instrument to a particular population for data collection. It is a necessary condition to achieve test validity. An instrument considered to be reliable, may not be valid. That is to say a reliable or unreliable test cannot automatically be valid for eliciting accurate measurement. This means that a test may be reliable for consistently measuring wrong attributes at repeated intervals without measuring what it is purported to measure (Adeleke, 2010).

Brogan (2009), in Igbacha (2011), affirms that reliability is established when a test is administered to a group with similar characteristics. Variance in the scores is usually attributed to true variance in the examinees' level of ability. The degree to which the two variances match is the estimated reliability in evaluation. Based on this theory, the accurate measure of a test is the ratio of the true and observed scores. This relates to the true score theory because if the true score equals the observed score, the error term must be equal to zero and the ratio of the true score to the observed score is a perfect 1.

A deviation from the ratio, whether positive or negative, is normally due to the strength of the error term.

Based on the above theory of reliability estimation, the question of whether the standardised BECE test items are reliable and valid for measuring the desired constructs cannot be ignored. Molenberghs and Verbeke (2005) stated that for normal distributed test scores, reliability of measurement is expressed in a statistical equation presented as:

$$\sigma_\theta^2 / \left(\sigma_\theta^2 + \sigma_\varepsilon^2 \right)$$

Where σ_θ^2 = the variance of the testee's ability (θ), and σ_ε^2 = the variance of assumed distribution error. It is generally interpreted as intra-class or group correlation. Directly using intra-class responses produces unobserved correlation with responses at a logit scale. Also, any reliability estimated through unobserved correlation is usually an unobserved trait. Molenberghs and Verbeke (2005) emphasised that the concern is basically directed towards observed score reliability and not latent scores. However, significant reliability estimates are on the unintended measures. For normal test scores distribution, unobserved and manifest correlations are conceited.

According to Briggs and Wilson (2007), reliability measure under IRT model estimation is done using manifest correlations which are not properly established and applied to dichotomous response unlike polytomous response. They further noted that this is always difficult to estimate because it involves the examination of integral test items without closed forms and are, therefore, not commonly estimated. They concluded that to address such a challenge, approximate reliability estimation is recommended for calibrating the reliability. Examples of approximate reliability are Cronbach's α, and the intra-group correlation. Under IRT frame work, reliability estimate is done using marginal error approach where the contribution of an individual item in the test is computed at average.

Cronbach and Shavelson, (2004) argued that applying Cronbach's α and Fisher's information measure have limited application of IRT model because such applications, under some conditions, yield negative coefficient, especially for standardized tests. Among the aims of reliability coefficient produced by a test is to produce standard index which can also constitute the validity of a given test. The reliability value produced by a test gives test developers the opportunity to get standard error of measurement (SEM) which enables test – takers to respond to the test items. Standard error of measurement is defined as the expected amount of variance of examinees' observed scores around their true scores. The most commonly used method of finding test reliability is although for Classical Test Theory (CTT), Kuder-Richardson = K-R20 or K-R21 which is normally used when items are not of the same difficulty level.

Formula: $K - R20 = \dfrac{k}{k-1} \left(1 - \dfrac{\sum pq}{s^2} \right)$

The above equation is normally used to dichotomously scored test items whose reliability coefficient range from -1.00 to 1.00 theoretically. In education, the emphasis is on the internal consistency measure of test items which is determined by the degree of

correlation within test items. The coefficient indicates the extent to which the test items jointly measure a particular construct. Its computation is equivalent to the ratio of true score variance to total score variance. Reliability coefficients of .80 are ranked very good, and that of .90 are ranked excellent. It is important to note that the closer a coefficient is to 1, the more reliable is the instrument.

Under IRT model, reliability determination of common interest using 1PL, 2PL or 3PLis the internal consistency of expected individual items and the expected sum of the scores. This involves the use of exact and approximate methods of reliability estimation. Approximate reliability estimation is used to determine reliability of a scale using observed scores for a dichotomous and polytomous response, employing either 1PL or 2PL models. These functions are normally gotten for both continuous data and dichotomous response test items. For binary data approximations such as Cronbach α, inter-class correlation and Fisher's information are employed (Molenberghs, Verbeke, & De Boeck, 2011).

Records have shown that students performed consistently poor in NECO BECE Social Studies and civic education presently known as National Values Education test. The percentage of students who scored between "Distinction and Credit (A and C)" grades in the subject from 2012 – 2016 was consistently below 50%. Over 50% of the students had between "Pass and Fail (P and F)" grades. Many factors can be ascribed to this low performance in Benue State. Principal among these is the quality of test items that produced such scores. It is doubtful whether NECO BECE had the desired properties of test items and persons statistics for the estimation of testees' latent trait. The poor performance could be because test items constructed by NECO BECE were biased to examinees' subgroups such as gender (male and female) and school location (rural and urban) Also, the level of internal consistency (reliability) of test items developed by NECO BECE was not ascertained. Based on the above development, this study is set to assess construct validity and reliability of NECO BECE 2017 in national values education as a determinant of test quality and student performance.

Research Questions

The following questions were developed and answered for the study:

1. Is NECO BECE National Values Education 2017 biased to testees in respect to:
   i. gender?
   ii. school location?
2. What is the Reliability estimate of NECO BECE National Values Education 2017 tests?

Methodology

The study is a non-experimental research of ex-post facto research design. It investigates BECE 2017 National Values Education for construct validity and reliability. It is a descriptive survey research because it involves a large sample population of JSS3 students without manipulating any variable. The study adopts multi-stage sampling method at various stages. At the first stage, purposive sampling technique was used to

select two senatorial districts (Zones B and C) for fair representation of the two major tribes in Benue State (Tiv and Idoma). Eight local governments (4 from each senatorial district) were randomly selected. At the second stage, proportionate to size technique was used to select 48 secondary schools (32 state public and 16 National BECE registered schools) across rural and urban settlements in the state. The third stage involved the selection of 1500 JSS3 students. An intact class was used from each selected school. National BECE 2017 National Values Education was adopted as the instrument with which data was elicited. The instrument consists of items with A – D options where the correct response is scored "1" and the incorrect "0". Data were collected through personal visits to the sampled schools by the researchers and research assistants. These were analysed using linear factor analysis, and exploratory structural equation modelling (ESEM). The reliability estimation was done through IRT second-order approach using R Software.

Result
The study result and findings are presented as answers to the following:
1. Is National BECE National Values Education 2017 biased to testees in respect gender?

Table 1
Model Result of National BECE 2-Factor ESEM with Gender as Covariate

Dim	Covariate	Est.	S.E	Est./S.E	Two-Tailed p-value	Remark
F1	ON Gender	0.025	0.028	0.887	0.375	Not Sig
F2	ON Gender	0.067*	0.029	2.326	0.020	Sig

Table 1shows the dimension assessment of NECO BECE National Values Education test using ESEM with gender as covariate to examine its direct effect on the factors' indicators. F1 and F2 in the table above refer to the 1st and 2nd factors underlying national BECE. The table shows that while gender has no significant effect on the first factor (F1, $p = 0.375$), it has a significant effect on the second (F2, $p = 0.020$) underlying NECO BECE test. It also shows that the NECO BECE 2017 test had no construct validity as one of its underlying factors functioned differentially with respect to the gender of the examinees.
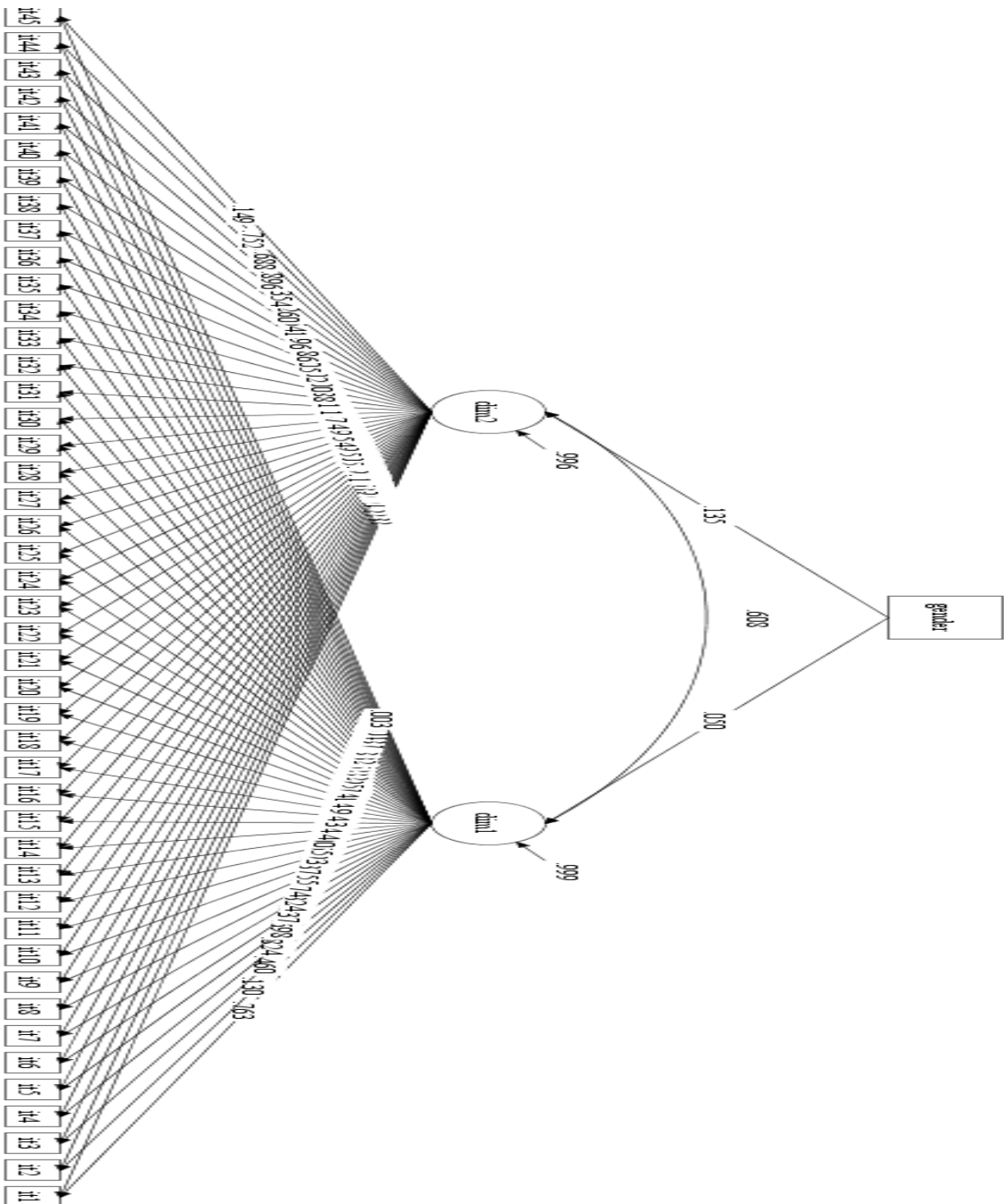
$x^2_{944} = 3090.081, p = 0.0000$; RMSEA $= 0.039$ (90% CI $= 0.037 - 0.040$, probability of RMSEA $\leq 0.05 = 1.000$). CFI $= 0.891$. TLI $= 0.880$

Figure 1: ESEM Analysis of NECO BECE NVET with Respect to Gender

Figure 1 presents the result of ESEM with gender as a covariate of one dimension or factor underlying the performance of the examinees in NECO BECE National Values Education. The figure shows that the two-factor model describing examinees' performance (construct validity) in NECO test, based on gender differences, is viable to some extent ($x^2_{944} = 3090.081, p = 0.0000$; RMSEA $= .039$ (90% CI $= .037 - .040$, with a probability of RMSEA $\leq .05 = 1.000$). CFI $= .891$. TLI $= .880$). This means, gender (male and female) affected the underlying trait measured by NECO BECE National Values Education test.

i. Is National BECE National Values Education 2017 biased to testees in respect of school location?

Table2

Model Result of NECO BECE 2-Factor ESEM with Respect to School Location as Covariate

Dim	Covariate	Est.	S.E	Est./S.E	Two-Tailed p-value	Remark
F1	ON School Location	0.018	0.028	0.652	0.514	Not Sig
F2	ON School Location	0.062*	0.028	2.169	0.030	Sig

Table 2shows the dimension assessment of NECO BECE 2017 National Values Education test using ESEM with school location as a covariate to the model to examine its direct effect on the factors' indicators. The table indicates that although school location has no significant effect on the first factor (F1, p = 0.514), it has a significant effect on the second factor (F2, p = .030) underlying NECO BECE test. It also shows that the NECO BECE 2017 test had no construct validity as one of its underlying factors functioned differently with respect to school location of the examinees.
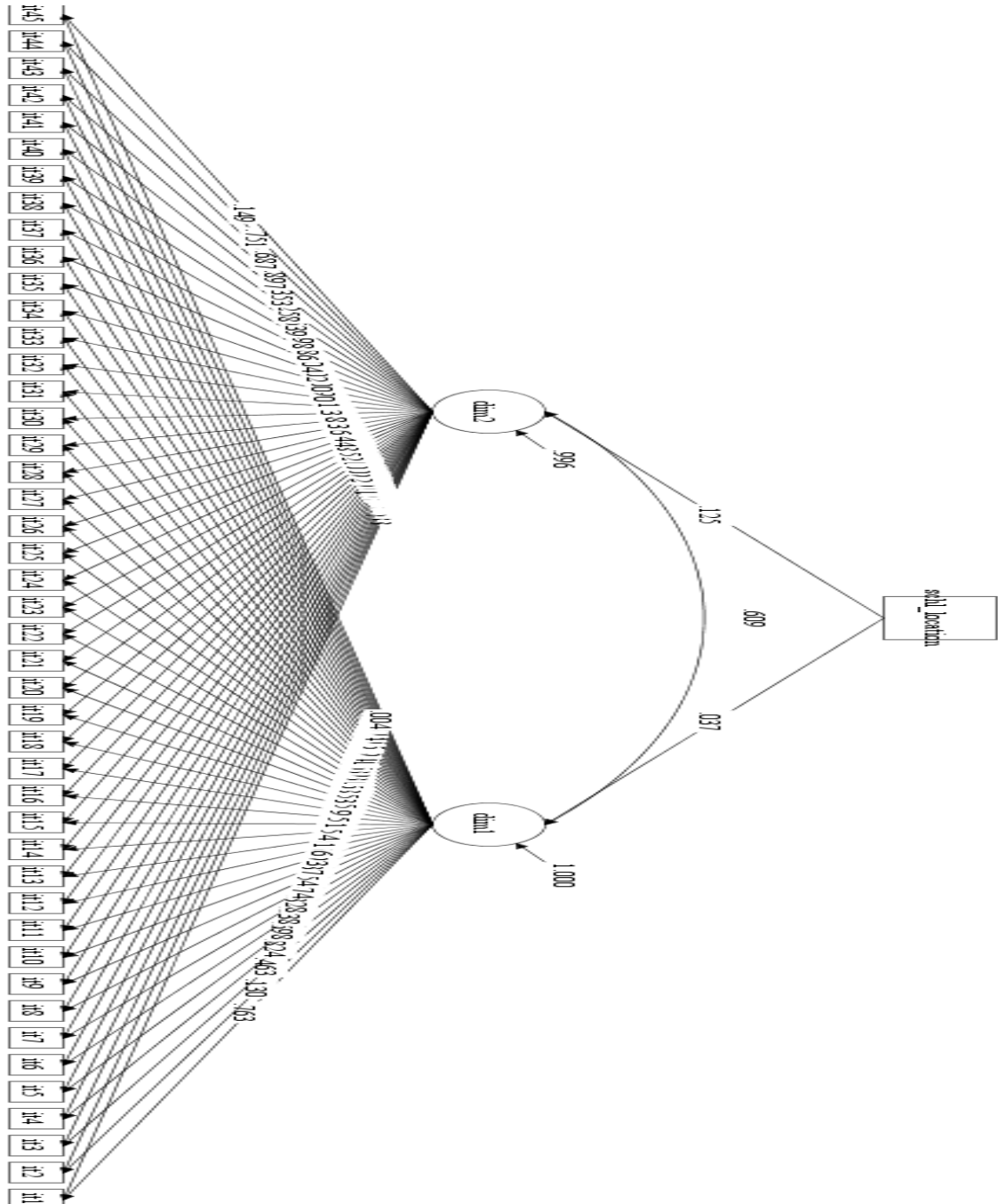
$x^2{}_{944} = 3116.183, p = 0.0000$; RMSEA $= 0.039$ (90% CI $= 0.038 - 0.041$, probability of RMSEA $\leq 0.05 = 1.000$). CFI $= 0.890$. TLI $= 0.879$

Figure 2:ESEM Analysis of NECO BECE NVET with Respect to School Location

Figure 2 presents the result of ESEM with covariate (school location) of one dimension found to underlie the performance of examinees in the NECO BECE National Values Education. The figure shows that the 2-factor model describing examinees' performance (construct validity) in NECO test items based on differences in school location was viable to some extent ($x^2{}_{944} = 3116.183, p = 0.0000$; RMSEA $= 0.039$ (90% CI $= 0.038$ - 0.041, with a probability of RMSEA $\leq 0.05 = 1.000$). CFI $= 0.890$. TLI $= 0.879$). This means, school location (rural and urban), as a subgroup, affected the underlying trait measured by NECO BECE National Values Education test.

2. What is the Reliability Estimate of National BECE National Values Education 2017 tests?

Table3
Marginal Reliability Estimate of NECO BECE National Values Education 2017

Test	Dimensions	Marginal reliability	Remark
NECO BECE		F1 = 0.8550695	High
	2	F2 = 0.6022473	Moderate

Table 3 shows marginal reliability estimation of NECO BECE 2017 National Values Education 2017 based on test dimensionality. The NECO BECE 2017 NVET with two underlying factors called 'socio-cultural values and national orientation' have marginal reliability coefficient of [(F1) r = 0.8550695 and (F2) r = 0.6022473]. This implies that only the first factor (F1) is highly reliable indicating high internal consistency of items.

Discussion of Findings
The study also discovered, as presented in Figure 1 and Table 1, that NECO 2017 NVE with two underlying factors lack construct validity in the assessment of students' ability in National Values Education. The first factor, called 'socio-cultural values' underlying NECO test, does not exhibit significant bias to male or female examinees in the area (F1, p = 0.375). However, the second factor called 'security consciousness', underlying the same NECO test, significantly exhibited bias to examinees based on gender (F2, p = 0.020). This means the items that measured security consciousness favoured one gender of students more than the other.

The findings on test bias to examinees' based on gender (construct validity) is in line with the findings of Goodness and Nneka (2015) that JSSCE 2009 Business Studies items in Rivers State discriminated against JSS3 students based on gender differences. The first factor (socio-cultural practice) underlying NECO test was not significantly biased to students based on their school location (urban and rural) differences (F1, p = 0.514). However, the second factor, 'security consciousness', exhibited significant bias against students based on school location (F2, p = 0.020). This finding is in line with Mokobi and Adedoyin (2014) who also found that Botswana Junior Certificate Examination Mathematics test was biased to students based on school location and gender differences. Nevertheless, the findings of this study on NECO BECE 2017 NVE

which shows that one of the underlying factors indicated bias and another indicated fairness to different examinees' subgroups confirmed the position of Sick (2010) that in cases of multiplicity of subgroup assessment, test items that differentiate by gender, age or any given subgroup does not make the test free of bias when other characteristics of examinees are considered for assessment of test construct validity. The findings also confirm Osadebe and Agbure (2018) who in their study on item differential functioning in social studies objective BECE 2014 in Delta Central Senatorial District discovered that the test significantly functioned differentially by gender, school location, school ownership, and socio-economic status. This implies that national (NECO) BECE 2017 test generally lacks construct validity to measure examinees' proficiency or ability for valid decision making.

This study's findings on reliability estimate revealed that NECOBECE 2017 NVE is reliable on two underlying factors. That the first factor, socio-cultural practices, is highly reliable implies that there was high internal consistency among items that l measured the factor. The reliability coefficient of the second factor (security consciousness) underlying NECO BECE 2017 NVE was moderately reliable. This also revealed that there was moderate internal consistency among items that measured the underlying factor. This made it evident that NECO BECE 2017 test items were reliable in measuring students' ability in National Values Education (NVE) but lack construct validity. This means that the test was consistently measuring the wrong thing among testees in Benue state. The finding agreed with Dimitrove (2003) and Anderson (2018) who stated that for binary items under IRT, the reliability of a test depends on the true-scores and IRT item parameters, and not necessarily the number of factors underlying the test. It can be inferred from the findings in Table 3 that the second factor underlying NECO BECE test with moderate reliability coefficient can hinder the consistency of test items when compared with the items measuring the first (F1) underlying factor with reliability coefficient. This could possibly account for students' poor performance in NECO or National BECE national Values Education in Benue state. This finding is similar to that of Bandele and Adewale (2013) who, in their investigation into comparative study of validity of WAEC, NECO and NABTEB 2012 mathematics tests, discovered that NECO BECE had low reliability coefficient (r = 0.512), WAEC (r = 0.815) and NABTEB (r = 0.752) at the significant level of 0.05, although, their study adopted CTT approach. Based on the purpose of BECE, which is for promotion and placement of students, the internal consistency of the items measuring the underlying factors should operate at the same level.

Conclusion
The findings of this study support the conclusion that National BECE NVE 2017 is reliable but not valid. It consistently failed to measure what it was purported to measure over the years. Therefore, students' low performance in the test is related to poor construct validity of the test, which is likely to affect the informed decision. The National

BECE items lack construct validity because their psychometric properties were not established using IRT approach during test development.

Recommendations

> National Examinations Council test developers should follow all the processes involved in test development during the construction of BECE test items particularly in national values education.

> Item Response Theory (IRT) approach should be used in establishing psychometric properties for public examination test items. The use of IRT for validity and reliability of test items enhances the quality and integrity of public examinations for decision making.

> NECO should validate BECE items and ensure that the test does not exhibit bias to any characteristics of the testees' subgroup such as gender, school location and age. The elimination of test bias to examinees' subgroups will enhance its construct validity and determine the extent to which the test result will be dependable. The Federal Ministry of Education should establish examination board committee or empower the existing ones to monitor the conduct of BECE test across the 36 state of the federation. The monitoring exercise should start from item development process implementation where test validity issues must be emphasized before test administration. This will entrench quality assurance in national BECE as a public examination.

References

Adeleke, J. O. (2010), *The Basics of Research and Evaluation Tools*. Some rest Ventures

Anderson, B. (2018).Item response theory reliability*:* Estimation of reliability coefficient for ability estimates and sum scores from IRT model package. *European Journal of   Social Science and Humanities*3, 41-52.

Bandele, S. O. & Adewale, A. E. (2013). Comparative analysis of the reliability and validity coefficients or WAEC, NECO and NABTEB constructed mathematics examinations. *Journal of Educational and Social Research* 3(2) 397 - 402

Briggs, D. C. & Wilson, M. (2007). Generalizability in item response modelling. *Journal of Education Measurement* 44, 131–155

Chima, C. A. (2014). Assessment of students' performance in basic education in Nigeria. *ICSHER, Journal* 3(2) 53 - 68

Cronbach, L. J., & Shavelson, R. J. (2004). My Current thoughts on coefficient alpha and success or procedures. *Educational and Psychological Measurement* 64, 391-418.

Culligan, B. (2005). Item response theory, reliability and standard error. *Aoyama Gakuin Women's Junior College*3, 48 - 66

DeAyala, R. J. (2009). *The Theory and Practice of Item Response Theory*. NY: The Guilford Press.

Dimitrove, D. M. (2003). Marginal true-score measurement and reliability for binary items as a function of their IRT parameters. *Applied Psychological Measurement*, 27(6) 440-458.

Goodness, O. & Nneka, C. A. (2015). Detecting Item bias in Rivers State JSSCE 2009 Business   Studies using IRT approach. *European Journal of Education and Development Psychology*, 4(1) 1 -11

Igbacha, S. A. (2011). *Testing Issues in Schools and Colleges*. Eggle Dove Press

Mokobi, T. & Adedoyin, O.O. (2014). Identifying location of baised item in 2010 Botswana Junior Certificate Examination Mathematics Paper One using the item response characteristics curves. *International Review of Social Sciences and Humanities,* 7(2) 63-82

Molenberghs, M. G. & Verbeke G. (2005).*Models for discrete longitudinal data*. Sprin

Molenberghs, M. G.,Verbeke, A.G. & De Boeck, P. (2011). Reliability measures in item response theory: manifest versus latent correlation functions. *Journal of Applied Statistics* 8(10), 1-21

Nenty, H. J. (2015). Conjugal relationship between research and measurement: A keynote address delivered at 1st EARNIA conference in Cameroon on 23rd – 27thNovember,   2015.

Nworgu, B.G. (2011). Differential item functioning: A critical issue in regional quality assurance Paper Presented at NERA Conference.

Ogbebor, U.C. (2012).Differential item functioning economics question paper of national examination council in Delta State, Nigeria. M.Ed Thesis, Institute of Education, University of Ibadan, Ibadan.

Onu, V.O.U. (2014), *Validation, standardization and calibration of mathematics and geography items for joint command secondary schools promotion examination in   Nigeria using item   response theory*. A Ph.D Thesis. Institute of Education, University   of Ibadan, Ibadan.

Osadebe, P. U & Agbure, B. (2018), Assessment of differential item functioning in Social Studies   multiple choice questions in basic education certificate examinations. *European Journal of Education Studies*, 4(9) 235-255.

Sideridis, G. D., Tsaousis, I. & Al-Sadaawi, A. (2018). Assessing construct validity of mathematics achievement: an application of multilevel structural equation modelling *(MSEM)*. *Frontiers in Psychology*,5(9) 14- 51

Sick, J. (2010), Assumptions and requirements of rasch measurement. *Shiken: JALT Testing and Evaluation SIG Newsletter,*14(2) 23–29.

Weiland, A., Durach, C. F., Kembro, J. & Treiblmairer, H. (2017). Statistical and judgmental criteria for scale purification. *Supply Chain Management* 22(4) https://www.researchgate.net/publication/

XinmingAn, & Yung, Y. F. (2014*). Item response theory: What it is and how you can use theIRT Procedure to Apply It*. SAS Institute Inc