JIRBDAI

Journal of Institutional Research, Big Data Analytics and Innovation

# INVESTIGATING THE ROBUSTNESS OF TEST SCORE EQUATING METHODS IN COMPARING AND MAINTAINING STANDARDS FOR STUDENTS' CONTINUOUS ASSESSMENT MEASURES IN SOUTH-SOUTH NIGERIA.

**UKO, MARY PATRICK[1], URSULA N. AKANWA[2], Eluwa, I. O.[3]**

[1]Department of Educational Foundations, School of education, Akwa Ibom State College of Education, Afaha Nsit, Akwa Ibom State, Nigeria. ORCID 0009-0000-6314-4376

[2&3] Department of Science Education, College of Education, Michael Okpara University of Agriculture, Umudike, Abia State, Nigeria.

Corresponding Author: okumma2000@yahoo.com

**ABSTRACT:** Continuous Assessment (CA) is a cornerstone of student evaluation in Nigerian secondary schools. However, discrepancies in CA practices raise concerns about fairness and standard maintenance.This study investigates the robustness of test score equating methods (linear equating, concurrent calibration,and separate calibration) in comparing and maintaining standards in CA measures across two South-South states in Nigeria. This study's scope was limited to South- South geopolitical zones to provides a starting point for understanding these challenges, but including more regions in further studies would make the findings applicable nationwide. The population of the study comprised of all Senior Secondary Three (SSIII) student's of 2021/2022 academic session in both states, and a sample of 1,605 students was drawn through multi-stage sampling procedure from the population. Parallel Mathematics Achievement Tests (MAT) with reliability of 0.84 and 0.88 respectively, were administered alongside existing mathematics CA scores. Three research questions and three hypotheses guided the study. Descriptive statistics was used in answering the research questions, while hypotheses were tested using independent t-test and Pearson product moment correlation coefficient. Scores were analyzed using linear equating, concurrent calibration, and separate calibration methods, leveraging SPSS and BILOG-MG. The findings revealed no significant differences in ability estimates using concurrent calibration but significant differences in linear equating, with Akwa Ibom outperforming Rivers (mean difference = 4.42). A moderate positive correlation (r = 0.42–0.43) between CA and Mathematics Achievement Test (MAT) scores was also observed.The study underscores the effectiveness of linear equating for standard maintenance, recommending its integration alongside standardized CA practices.To enhance the generalizability of findings, future studies should include more regions. Furthermore, contextualizing global equating practices to Nigeria's educational challenges and exploring advanced equating methods like equipercentile equating are recommended.

**KEYWORDS:** Equating design, robustness, test scores, standard and maintenance.

## I. INTRODUCTION

The role that assessment plays in the educational development of a person and the nation is very significant. It is used for quality control, standard maintenance and a means of determining accountability level displayed by stakeholders, determination of teaching /learning effectiveness as well as finding out student achievement. Assessment that is regular, cumulative and comprehensive is said to be a continuous assessment. Continuous Assessment (CA) is essential for evaluating student learning and maintaining educational standards.However, inconsistencies in CA practices across Nigerian states lead to misinterpretations of scores, disparities in grading, and challenges in ensuring fairness and standard maintenance. For example, CA scores make up 30–60% of terminal assessment scores, yet differences in

scoring practices undermine their reliability. Continuous assessment practice in the Nigerian context, have the teachers developing different tests and using them to determine the ability, proficiency or curriculum related achievements of students. A test is a set of structured questions in which individuals are expected to give answers to and each question in the test has a preferred answer (Evans, Uko & Ekim, 2022). The authors further noted that the response of the individual quantifies the behaviour of that individual. Agah, 2013 also described a test as an instrument which can be utilized in detecting some qualities, traits, characteristics, attributes, etc possessed by a person, an object or a thing. One major goal of testing is to reveal the latent ability of learners, which is determined by the number of correct responses made by the test taker and reported as raw test score or some norm-based transformation of it. The transformation of raw scores to standardized scores through any means of transformation like scaling or equating process, has the resulting scores referred to as measures (Udofia and Uko, 2016). The statistical characteristics of the tests used for continuous assessment in Nigeria vary across schools and depends on the characteristics of the population they were designed for. The characteristics of the schools and the teachers in these schools also vary. Based on these variations, measurement errors are bound to occur from tests developed by individual teachers and the scores generated from them. Scores obtained from different tests or examinations are often added up by teacher to get an individual student's total score for relative positions of students in the class. This act of adding up raw scores may lead to misinterpretation of marks because each assessment tool is crafted for a specific purpose and may not have the same mean and standard deviation. Therefore, for any comparison to be made, their test score should be standardized and maintained through an appropriate method.

Assessment is viewed from different perspectives by different teachers and authors. Assessment according to Ajuonuma (as cited in Ogunleye and Omolayo, 2016) and Emaikwu (2014) is a process of gathering data and fashioning them into interpretable form for decision-making. It involves the collection of data with the view to making valued judgment about the quality of a person, object or event, for measuring behaviours and taking relevant decisions about students, curriculum and instruction as well as educational policy. Assessment system, the standard on which it is based on and all aspect of the assessment should treat students equally and be sensitive to all aspects of school experience both within and outside the classroom which includes cultural, ethnical, class and gender differences, and to disabilities, and must be valid for and not penalize any group of learners (Faleye and Adefisoye, 2016; Uko, 2021). To ensure fairness uniformity and equity, students should have multiple opportunities to meet standards in different ways and not depending upon a single test score. It therefore follows that several forms of tests should be used to obtain students' final test scores. This is exactly what Continuous Assessment (CA) practice in Nigeria is intended to achieve. However, the problem is in the implementation and this problem arises probably because there are no nationally standardized tests used for student's continuous assessment (CA). Rather each teacher in every school in the country prepares their own Continuous Assessment Test (CAT). (National Teachers Institute, 2015). In the presentation on assuring fairness in the continuous assessment component of school based assessment practice in Nigeria, Nwoke, Osuji and Agi (2017) argued that, the entire practice of CA is faced with laxity and they are; laxity in timing, mode that the CA exercise takes, and in relation to the content of CA, and that continuous assessment scores (CAS) has generated a lot of controversies between schools and examination bodies like West African Examinations Council (WAEC) and National Examinations Council (NECO). Schools are blaming examination bodies of not using the CAS forwarded to them while the examination bodies are in turn accusing the schools of inflating the CAS sent to them by the school. There is no uniformity in the continuous assessment scores (CAS) incorporated into the terminal assessment scores in the different states. The CAS stands as, 30% in Cross River, 60% in Rivers states, 40% in Enugu, Akwa Ibom, Anambra and Lagos states, of every examination at the secondary school level.

A critical look at the CA scores assigned to students by schools in Cross River, Akwa Ibom state and River's state shows a wide difference in the scores. This attest to the fact that, there is no uniformity in the practices of CA, this may have caused some lapses in the grading, interpreting and reporting of students' performance and hence the need to device a means of making the CA scores comparable (Afemikhe, 2007). The CAS forwarded to WAEC and NECO in all subjects and Mathematics in particular indicate that most students

perform above average. But the failure rate of students in public examinations is still high. In comparability research carried out by Makiney, Rosen, and Davis, (2003) emphasis was on the differences in the means and standard deviations of test scores. Raju, Laffitte, and Byrne (2002) state that "without measurement equivalence, it is difficult to interpret observed mean score differences meaningfully." It is therefore necessary not to only use mean and standard deviation in comparing student CA in Akwa Ibom and Rivers states but to include other measurement issues such as item discrimination, difficulty and guessing parameters. Currently, examination bodies like NECO use T-score to transform CAS before incorporating into TAS. This approach does not consider how scores are obtained, does not ensure that scores emanating from CA are authentic and actual reflection of the students' academic achievement. Based on the above context, enhancement strategies such as moderation, self assessment and test scores equating have been suggested as educational standards control mechanism in Nigeria (Afemikhe, 2007). Moderation and self-assessment have been practiced mostly at the tertiary level and not at the secondary level of education in Nigeria, whereas test score equating is not practiced at all. The statistical process of making test scores comparable is called test equating (Kolen and Brennan, 2004). Arai and Mayekawa (2011) described test equating as a statistical procedure of ensuring that candidate are measured against the same criterion-referenced standard regardless of the test administration. Test equating accounts for variations in the difficulty and other statistical characteristics of different tests so that scores from equated tests have comparable meaning (Chong and Sharon, 2005; Ofqual, 2010; Vin der Linden, 2006). Based on the above definitions, one can deduce that, test equating is aimed at placing the scores obtained from different forms of a test on a benchmark scale.

In conducting test score equating, numerous methods are available and are anchored mainly on the framework of two test theories namely; classical test theory (CTT) and item response theory (IRT). Classical test theory is founded on a test score model which assumes that, there are no perfect measures of ability. Each examinee's observed score (X) is comprised of True Score (T) and error score (E) (Schumacker, 2005). Item Response Theory (IRT) is based on the assumption that, there is a mathematical function that describes the relationship between an examinee proficiency and probability that an examinee will answer an item correctly (Chong, 2007). The logistic model which dictated the three parameters: the item discrimination parameter (a), the item difficulty parameter (b), and the lower asymptote or guessing parameter (c) is known as the three-parameter logistic model. The ultimate goal of both Classical Test Theory (CTT) and item response theory (IRT) is to establish the position of the individual along some latent dimension (ability). This study specifically examined linear equating method and concurrent calibration of IRT. These methods were chosen because of their appropriateness in handling the problem under investigation. Linear equating ensures that, the means and standard deviations of two test forms for a particular group of examinees are set to be equal, setting the mean of the examinees' ability levels to zero and the standard deviation to one. This method is mostly appropriate if the groups taking the test forms are equivalent or have equal ability, but can also be appropriate in a non-equivalent anchor test group design (Kolen and Brennan, 2004). Three parameter logistic model (3PLM) concurrent calibrations combined two or more test forms that share common items into a single data set and then calibrated simultaneously. The process of calibrating automatically equates the test forms (Morrison and Filzpatrick, cited in Agah, 2013). The goal of equating is for scores on multiple test forms to be used interchangeably. Test scores equating also ensure that no examinee is disadvantaged or advantaged because of the form of test taken. This study aims to evaluate the robustness of test score equating methods for comparing and maintaining standards in CA scores across South-South Nigeria. Standardizing CA scores is crucial to ensure equity in student performance evaluations.Methods like test equating have been successful globally, but their application in Nigerian secondary schools is underexplored. Test equating, a statistical process, ensures that scores from different test forms are comparable. Using frameworks like Classical Test Theory (CTT) and Item Response Theory (IRT), equating addresses disparities in test difficulty and other characteristics. Globally, test equating practices ensure fairness, reliability, and standardization in student assessments, particularly in large-scale testing programs.For example, methods like linear equating and equipercentile equating have been used effectively in standardized testing systems to address score comparability issues and maintain assessment

integrity.However, Nigeria's educational system presents unique challenges. The lack of standardized CA tests across schools leads to significant variability in how scores are assigned, often influenced by teacher biases and resource limitations.Moreover, some states inflate CA scores, further complicating efforts to maintain standards. In this context, global equating methods provide a practical solution.Linear equating, for instance, can standardize scores by adjusting them to a common scale, enabling fair comparisons across states with different CA practices.Similarly, equipercentile equating can account for nonlinear differences in scoring patterns, making it particularly useful in heterogeneous educational systems like Nigeria's. Implementing these methods within the Nigerian context would require robust training programs for teachers and policymakers, as well as investments in standardized test development and equating software.By contextualizing these global practices to Nigeria's specific challenges, the educational system can achieve greater equity and reliability in student assessments.

## II. LITERATURE REVIEW

### TEST SCORE EQUATING METHODS

Test score equating is a statistical process that ensures scores from different test forms are comparable, maintaining fairness and validity in assessments.This process is crucial when multiple test forms are administered to account for variations in difficulty levels.Equating allows scores to be used interchangeably, facilitating accurate interpretations of student performance. Several equating methods are prevalent: Linear Equating: Adjusts scores based on the mean and standard deviation, assuming a linear relationship between different test forms.This method is straightforward but may not account for all variations in test difficulty. Equipercentile Equating: Matches percentiles of score distributions from different test forms, allowing for nonlinear relationships.This method is more flexible but requires larger sample sizes for accuracy. Item Response Theory (IRT) Equating: Utilizes models that consider the probability of a correct response based on item characteristics and examinee ability.IRT provides a robust framework for equating, especially when test forms vary significantly in content and difficulty. Recent studies emphasize the importance of equating in maintaining the validity and reliability of assessments.For instance, equating is essential in standardized testing programs to ensure that scores are comparable across different administrations, thereby upholding the integrity of the testing process.

### CONTINUOUS ASSESSMENT PRACTICES IN NIGERIAN SECONDARY SCHOOLS

Continuous assessment (CA) was introduced in Nigeria to provide a more comprehensive evaluation of students' abilities, encompassing cognitive, affective, and psychomotor domains. The National Policy on Education advocates for CA as a means to reduce the overemphasis on one-time examinations and to promote continuous feedback in the learning process. However, the implementation of CA in Nigerian schools faces several challenges: Inconsistencies in Implementation: Studies have found that CA practices vary widely among schools, with some teachers lacking a clear understanding of CA procedures.This inconsistency leads to unreliable assessments of student performance. Teacher Preparedness: Many teachers have not received adequate training in CA methodologies, resulting in improper implementation and record-keeping.This gap undermines the effectiveness of CA in achieving its intended educational outcomes. Resource Constraints: Limited resources and large class sizes hinder the effective practice of CA, as teachers are unable to provide individualized assessments and feedback.This situation compromises the quality of education and the accuracy of student evaluations. To address these challenges, researchers recommend: Standardization of CA Practices: Developing clear guidelines and training programs for teachers can promote uniformity in CA implementation across schools.Standardization ensures that all students are assessed fairly and accurately. Integration of Equating Methods: Applying test equating methods to CA scores can enhance comparability and maintain standards across different schools and regions.This integration ensures that CA scores are reliable indicators of student performance. Policy Reforms: Educational policymakers should consider reforms that address the systemic issues affecting CA implementation, such as providing adequate resources and reducing class sizes.Such reforms are essential for creating an environment conducive to effective continuous assessment. In conclusion, while continuous

assessment holds promise for enhancing educational outcomes in Nigeria, its effectiveness is currently hampered by implementation challenges.Incorporating robust test equating methods and addressing systemic issues can improve the reliability and validity of CA practices, leading to better educational standards nationwide.

## TEST SCORE EQUATING METHODS

Test score equating is essential for ensuring fairness and validity in educational assessments (Kolen & Brennan, 2004).Linear equating adjusts test scores based on mean and standard deviations, making it suitable when test groups have equivalent abilities (Chong & Sharon, 2005). Meanwhile, equipercentile equating allows for nonlinear score adjustments, addressing differences in score distributions across tests (Von Davier & Kong, 2005). Item Response Theory (IRT) provides a robust framework for equating, focusing on item characteristics such as difficulty, discrimination, and guessing parameters (Arai & Mayekawa, 2011).IRT-based methods, such as concurrent calibration, effectively align scores across diverse test forms by calibrating shared anchor items simultaneously (Van der Linden, 2006). Recent advancements in equating methods underscore their importance in standardizing assessments, particularly in contexts where multiple test forms are required for large-scale assessments (Makiney et al., 2003; Ofqual, 2010).

## CONTINUOUS ASSESSMENT PRACTICES IN NIGERIA

Continuous assessment (CA) in Nigeria was introduced to provide a holistic evaluation of students, incorporating cognitive, affective, and psychomotor domains (National Teachers Institute [NTI], 2015).However, its implementation is fraught with challenges.Studies highlight inconsistencies in CA practices, with variations in teachers' understanding and application of assessment techniques leading to disparities in student evaluations (Faleye & Adefisoye, 2016; Ogunleye & Omolayo,2016). Teacher preparedness is a critical factor, with many teachers lacking adequate training in CA methodologies (Emaikwu, 2014).Additionally, resource constraints, such as large class sizes and limited infrastructure, hinder effective implementation (Ayodele & Balogun, 2017). Efforts to standardize CA practices include introducing equating methods to ensure comparability across regions (Afemikhe, 2007).IRT and Classical Test Theory (CTT) are increasingly recognized for their potential to improve the accuracy and fairness of CA evaluations (Schumacker, 2005; Raju et al., 2002).

Von Davier and Kong (2005) carried out a study that examined the application of linear and concurrent calibration equating methods in large-scale assessments.Results showed that concurrent calibration is highly effective when test forms share anchor items, yielding no significant differences in ability estimates between groups.The researchers concluded that concurrent calibration aligns test forms more accurately by minimizing statistical errors in score comparisons.This supports the expectation of no significant differences in ability estimates when concurrent calibration is applied. The study conducted Arai and Mayekawa (2011), compared equating methods (concurrent calibration and separate calibration) using IRT frameworks.Findings revealed that concurrent calibration effectively handles variations in test forms by simultaneously calibrating anchor items, leading to negligible differences in ability estimates.Separate calibration, however, required additional linking steps, which introduced small but non-significant differences. The results validate the hypothesis that concurrent calibration reduces discrepancies in ability estimates.

Makiney, Rosen, and Davis (2003) conducted a study, this research focused on the linear equating of test scores across states using standardized achievement tests.The findings showed significant differences in scores between states when equated using linear methods, attributed to differences in population ability distributions and test difficulty.The authors recommend linear equating only for comparable groups. Linear equating may yield significant differences in ability estimates across heterogeneous populations, suggesting the rejection of the hypothesis. Kolen and Brennan (2004): In their extensive analysis of equating methods, the authors demonstrated that linear equating assumes equivalent ability distributions across groups.When applied to non-equivalent groups, it often produces significant differences in mean scores.This method is

best suited for homogeneous groups or small-scale assessments.The findings reinforce the expectation of significant differences in equated scores when groups are heterogeneous, challenging the null hypothesis.
A study was conducted by Ayodele and Balogun (2017). This study examined the relationship between continuous assessment (CA) and standardized test scores in mathematics.Using Pearson's correlation, results revealed a moderate positive correlation (r = 0.45), suggesting that CA scores moderately predict performance in standardized tests.The authors attribute this relationship to differences in scoring standards and teacher biases in CA practices. Faleye and Adefisoye (2016) carried out a study investigating the link between continuous assessment and terminal examination scores in Nigerian schools, the study found a strong positive relationship (r = 0.63) in mathematics. However, discrepancies in scoring standards and non-standardized CA practices weakened the predictive accuracy in some schools.The authors recommend standardized CA practices to improve reliability. The findings highlight a significant relationship between CA and standardized test scores, suggesting the null hypothesis is untenable.

### ADDRESSING THE CHALLENGES

This study situates its analysis within the broader landscape of test equating methods, emphasizing the global significance of linear, equipercentile, and IRT-based approaches. However, the unique challenges of Nigeria's CA system, including the lack of standardized assessments and resource constraints, require deeper contextualization.For instance, inconsistencies in teacher training and scoring methods contribute to the disparity in CA outcomes.Integrating global best practices with these local challenges can bridge gaps in equity and reliability. To address CA's challenges, researchers recommend policy reforms, such as the provision of resources and teacher training, to ensure uniformity in assessment practices (Nwoke et al.,2017). Additionally, integrating equating methods into CA practices can standardize scores, enabling fair comparisons across schools and states (Kolen & Brennan, 2004; Chong & Sharon, 2005).
This study hypothesizes that robust test equating methods, particularly linear equating and concurrent calibration, effectively standardize CA scores for comparability and standard maintenance. Thus, the inherent problem in the derivation of CAS and comparability of CAS for standard maintenance among schools or states are the gap this study intends to fill. Despite advancements in test equating methods, their application in Nigerian CA systems remains limited.The lack of standardized tests and reliance on teacher-generated assessments exacerbate score disparities. This study addresses these gaps by applying linear equating and concurrent calibration methods to standardize CA scores in Nigerian secondary schools, providing foundation for equitable assessment practices. The researchers therefore designed this study to examine the robustness of test scores equating methods in the comparison and standard maintenance of students' continuous assessment measures in secondary schools in South-South, Zone, Nigeria. Specifically, the study:

(a). Determine the mean ability estimates of students in MAT in states one and two when their scores are equated through concurrent and separate calibration
(b). Determine the mean ability estimates of students scores scaled through linear equating in states one and two.
(c). Determine the relationship between the performances of students in mathematics continuous assessment tests and mathematics achievement tests.

## III. METHODOLOGY

The research employed a non-equivalent anchor test (NEAT) group design, which is appropriate for comparing scores from different groups while using shared anchor items to standardize the measurement scales. The research was carried out in South-South Zone, Nigeria. Addressing limitations of generalizability, the study acknowledges its limitation of being confined to Akwa Ibom and Rivers states, which may restrict the generalizability of its findings. Future studies should include states from other geopolitical zones (e.g., North-West, South-East) to enhance the applicability of findings across the nation and provide a more comprehensive understanding of Continuous Assessment (CA) practices in Nigeria. The following three research questions guided the study: 1. Research question one examines whether the ability

estimates derived using concurrent calibration (a method within item response theory) show equivalence between the two states. 2. This question focuses on the use of linear equating to determine differences or similarities in mean ability estimates between the two states, emphasizing its sensitivity to disparities in CA practices. 3. This question seeks to determine whether CA scores can reliably predict students' performance in standardized achievement tests, using correlation analysis. The study tested the following null hypotheses ($H_0$) at a 0.05 level of significance: $H_{01}$: This hypothesis evaluates whether concurrent calibration can establish statistical equivalence in ability estimates, accounting for test characteristics across states. $H_{02}$:This hypothesis tests whether linear equating detects significant differences in ability estimates between the two states. $H_{03}$: This hypothesis assesses whether a moderate or strong positive correlation exists between CA and MAT scores.

The sample size of 1,605 respondents used for this study was selected from a population of 30,967 students in 8 secondary schools out of 456 public secondary schools in the study area. Cluster sampling technique was used for the selection of 4 secondary schools each in Akwa Ibom and Rivers States. This study's scope was limited to South- South geopolitical zones to provides a starting point for understanding these challenges, but including more regions in further studies would make the findings applicable nationwide. To acquire relevant data for the study, two parallel 40 items multiple choice Mathematics Achievement Test (MAT) developed by the researchers was used in equating students' continuous assessment scores in mathematics from both states. Twenty (20) items are common to both tests while the other 20 items are unique to each test and was large enough to enhance high precision in the determination of the psychometric properties of the items as well as producing good ability estimates of the students. The reliability of the instruments was established using Kuder-Richardson 20 and the result showed a reliability of 0.86 and 0.84; thus, making the instruments good enough to justify its use in the study. The instruments were administered to the students in their various schools and states by the researchers and some research assistants who were mathematics teachers in the sampled schools. The items were scored according to scoring key which was 1 for correct option and 0 for incorrect option. The MCA scores of each student was recorded on the script by their mathematics teacher. Addressing Research Question 1 and $H_{01}$: Concurrent calibration was used to estimate the mean ability levels of students in the two states. The estimates were compared using descriptive statistics (means, standard deviations and scores difference that matter (SDTM)) and an independent t-test to determine equivalence. Addressing Research Question 2 and $H_{02}$: Linear equating was applied to CA and Mathematics Achievement Test (MAT) scores to scale them onto a common metric.The resulting mean scores were analyzed using descriptive statistics and tested for significant differences using an independent t-test. Addressing Research Question 3 and $H_{03}$: The relationship between CA and MAT scores was evaluated using Pearson's product-moment correlation coefficient (PPMC).The correlation results were used to determine the predictive reliability of CA scores in assessing student performance. The chosen methods ensure comparability across varying test forms and address disparities in scoring practices, making them suitable for this study's objectives. Each result then indicated whether the hypothesis should be accepted or rejected. For the null hypothesis, the decision rule was to reject the null hypothesis, if the calculated r is greater than the critical r at the .05 level of significance or accepted if the other way round.

## RESEARCH QUESTIONS

(a). What is the student's mean ability estimates in state One and Two when equating their scores through concurrent and separate calibration?
(b). What are the students mean estimates scores when scaled through linear equating in state One and Two?
(c). Is there any relationship between the performance of students in mathematics continuous assessment tests and mathematics achievement tests?

### HYPOTHESES

**H01:** There is no significant mean difference in the students' ability estimates for scores equated through concurrent and separate calibration in states One and Two.

**H02:** There is no significant mean difference in the students' ability estimates for scores equated through linear equating in state One and Two

**H03:** There is no significant relationship between students' performances in continuous assessment test and mathematics achievement test.

## IV. RESULTS AND DISCUSSION

**RESEARCH QUESTION 1:** What are the students mean ability estimates in state One and Two when equating their scores through concurrent and separate calibration?

**Table 1:** Students mean ability estimates in state One and Two for scores equated through concurrent calibration.

| State | $\bar{X}$ | SD | $S^2$ | RMS | RMS($S^2$) |
|---|---|---|---|---|---|
| One | 1.82 | 0.42 | 0.18 | 0.71 | 0.56 |
| Two | 1.79 | 0.61 | 0.37 | 0.66 | 0.48 |
| Mean Diff. | 0.03 | | | | |

*Akwa Ibom is state one and Rivers is state two.*

From Table 1, the mean ability estimates of students in state One is 1.82 and that of state Two is 1.79. The Table shows also that, the standard deviation, variance, root mean square and root mean square variance posteriori of ability estimate for the two states are; State One [SD = 0.42, $S^2$ = 0.17, RMS= 0.71 RMS($S^2$) = 0.56] and State Two [SD = 0.61, $S^2$ = 0.37, RMS= 0.66 RMS($S^2$) = 0.48] The concurrent calibration appears to show some slight difference in the students' mean ability estimates in state One and Two. However, the score difference that matter (SDTM), did not show any difference in mean and other moments.

### HYPOTHESIS 1

HO1: There is no significant difference in the students' ability estimates for scores equated through concurrent calibration in states One and Two.

This hypothesis was tested at 0.05 level of significance using the summary of the theta (θ ) values from concurrent calibration

**Table 2:** Independent T-test Analysis of Students' Performance when Scores are Standardized through Concurrent Calibration Equating.

| **State** | **N** | $\bar{X}$ | **SD** | **df** | **t-cal** | **t-crit** | **Decision** |
|---|---|---|---|---|---|---|---|
| ONE | 853 | 1.82 | 0.42 | 1603 | 1.22 | 1.96 | NS |
| TWO | 752 | 1.79 | 0.61 | | | | |

α = 0.05, NS= non-significant

The distribution moments (mean and standard deviation) of concurrent calibration equating were used in testing this hypothesis. The results of the study showed no significant difference in ability estimates between Akwa Ibom and Rivers states when concurrent calibration was applied ($t = 1.22$, $p > 0.05$), hence, the null hypothesis was upheld. Therefore, there is no significant difference in the students' ability estimates for scores equated through concurrent and separate calibrations in states One and Two.

**RESEARCH QUESTION 2:** What are the students mean estimates scores when scaled through linear equating in state One and Two?

To answer this research question, the students' scores in both MCA and MAT were first converted to 100%, to bring all the scores from different state and test forms to a common denominator. The first two distribution moments (mean and standard deviation) in MCA and MAT for the two states were then used to standardize the scores using linear equating to answer this research question and the summary result is presented in Table 3 below.

**TABLE 3:** Students mean estimate score scaled using linear equating

| State | $\bar{X}$ | SD | $S^2$ | S. E |
|---|---|---|---|---|
| **One** | 56.07 | 7.53 | 56.70 | 0.21 |
| **Two** | 51.65 | 7.13 | 50.83 | 0.21 |
| **Mean Diff.** | 4.42 | | | |

Table 3 revealed that, the students' mean performance when standardized using linear equating method are 56.07 and 51.65 for state One and state Two respectively. The mean difference of the scaled score between the two states is 4.42 in favour of state One. This mean difference indicates that, students in state One have higher ability estimate than those in state Two when MCA scores and MAT scores are used in determining their ability.

**Ho 2:** There is no significant mean difference in the students' ability estimates for scores equated through linear equating in state One and Two.

Measures from linear equating were further subjected to statistical analysis using SPSS to test this hypothesis and the result obtained is shown in Table 4.

**TABLE 4:** Independent T-test Analysis of Students' ability Estimates when Scores Standardized through Linear Equating.

| State | N | $\bar{X}$ | SD | df | t | Sig | Decision |
|---|---|---|---|---|---|---|---|
| **One** | 853 | 56.07 | 7.53 | 1603 | 13.28 | 0.00 | Sig. |
| **TWO** | 752 | 51.65 | 7.13 | | | | |

α = 0.05, S= Significant

Table 4 showed that t-calculated was 13.28 with associated probability value of 0.00. Since the associated probability (0.00) is less than 0.05, the null hypothesis was not accepted. Therefore, there was a significant difference in the students' ability estimates for scores equated using linear equating in state One and Two.

**RESEARCH QUESTION 3:** What is the relationship between the performance of students in mathematics continuous assessment tests and mathematics achievement tests?

**Table 5:** Pearson's Product Moment Correlation Analysis of MCA and MAT

| State | Test | N | $\bar{X}$ | SD | r | Sig | Decision |
|---|---|---|---|---|---|---|---|
| **One** | MCA | 853 | 56.07 | 7.53 | 0.42 | 0.00 | S |
| | MAT | 752 | 35.34 | 15.68 | | | |
| **TWO** | MCA | 853 | 51.65 | 7.13 | 0.43 | 0.00 | S |
| | **MAT** | **753** | **22.53** | **14.90** | | | |

*α = 0.05, S=Significant*

For this research question to be answered, MCA scores and MAT scores were first converted to 100%. The scores from MCA and MAT in state One and Two were correlated using Pearson's product moment correlation and the coefficient of 0.42 and 0.43 respectively were obtained. The result shown on table 5 and the coefficients obtained show a moderate positive relationship between MCA and MAT in both states.

**HO 3:** There is no significant relationship between students' performances in continuous assessment test and mathematics achievement test.

The result presented in Table 5 showed that, there exit a moderate positive relationship between MCA scores and MAT scores, state One was 0.42 and state Two 0.43 with associated probability of 0.00 for both. Since the associated probability in each case was less than 0.05 level of significance with 851 and 750 degrees of freedom respectively, it means that the result was significant and the null hypothesis was not upheld. This visual supports the findings related to Research Questions 1 and 2 and Hypotheses H01 and H02. Figure 1 below compares the mean ability estimates for Akwa Ibom and Rivers states, derived through concurrent calibration and linear equating. The graph illustrates whether significant differences exist between the methods and the two states.

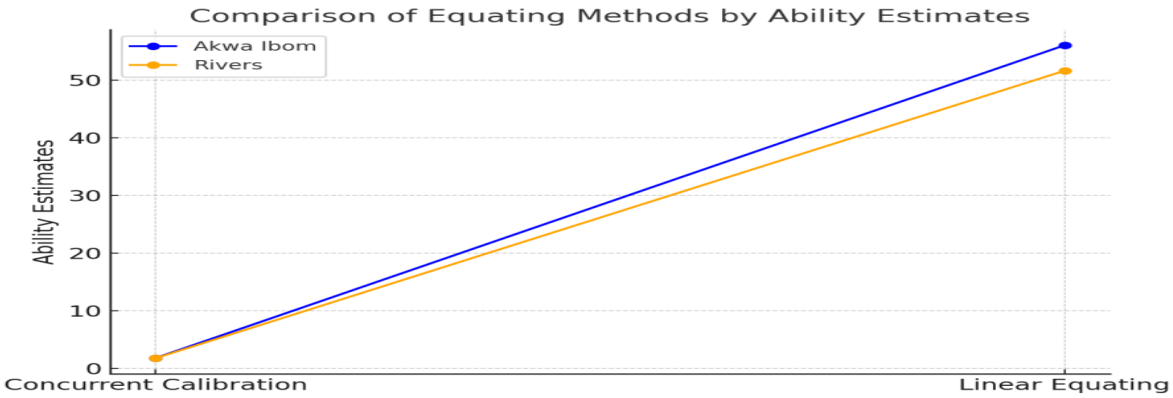


**Figure 1:** Comparison of Equating Methods by Ability Estimates

As seen in Figure 1, the mean ability estimates for Akwa Ibom and Rivers are slightly different for both equating methods, suggesting that linear equating slightly favors one state over the other. The results showed no significant difference in ability estimates between Akwa Ibom and Rivers states when concurrent calibration was applied ($t = 1.22$, $p > 0.05$). However, linear equating revealed significant disparities, with Akwa Ibom students outperforming Rivers by a mean difference of 4.42 ($t = 13.28$, $p < 0.05$). This indicates that linear equating is more sensitive to regional disparities, making it a valuable tool for standardizing CA scores. These findings provide insight into the robustness of the equating methods across different regional contexts.

This Bar chart supports the findings related to Research Questions 2 and Hypothesis H02. The Bar chart below compares the mean Standardized Scores in Akwa Ibom and Rivers States Using Linear Equating. The bar chart illustrates whether significant differences exist between the mean Standardized Scores of the two states.

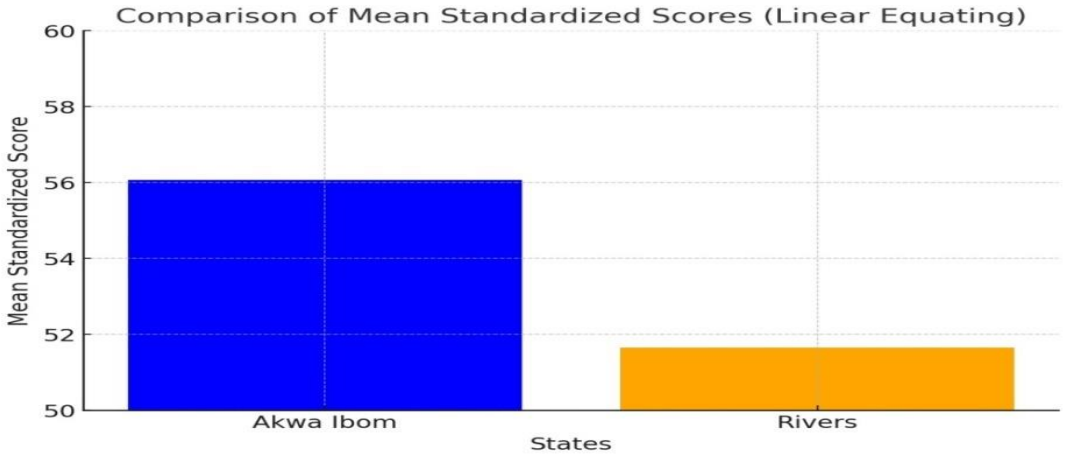


**Figure 2:** Bar Chart showing Comparison of Mean Standardized Scores in both States Using Linear Equating.

The Bar Chart shows Mean Standardized Scores Using Linear Equating. This chart compares the mean scores for Akwa Ibom (56.07) and Rivers (51.65), highlighting the significant mean difference of 4.42 in favor of Akwa Ibom. It aligns directly with the hypothesis test result, which rejected $H_{02}$ ($p < 0.05$). This significant disparity ($t = 13.28$, $p < 0.05$) indicates that linear equating highlights regional differences in Continuous Assessment (CA) scores. The bar chart emphasizes the mean differences between states, and simplifies the statistical comparison by representing the mean scores visually, making it easier for non-technical audiences to interpret the findings.

This scatterplot and line graph are highly relevant to Research Question 3 and Hypothesis H03, as it directly visualizes the correlation between Continuous Assessment (CA) scores and Mathematics Achievement Test (MAT) scores. Figure 2 and 3 below demonstrates the relationship between Continuous Assessment (CA) scores and Mathematics Achievement Test (MAT) scores for Akwa Ibom and Rivers states.The visuals highlights potential differences in the strength of correlation between the two states.

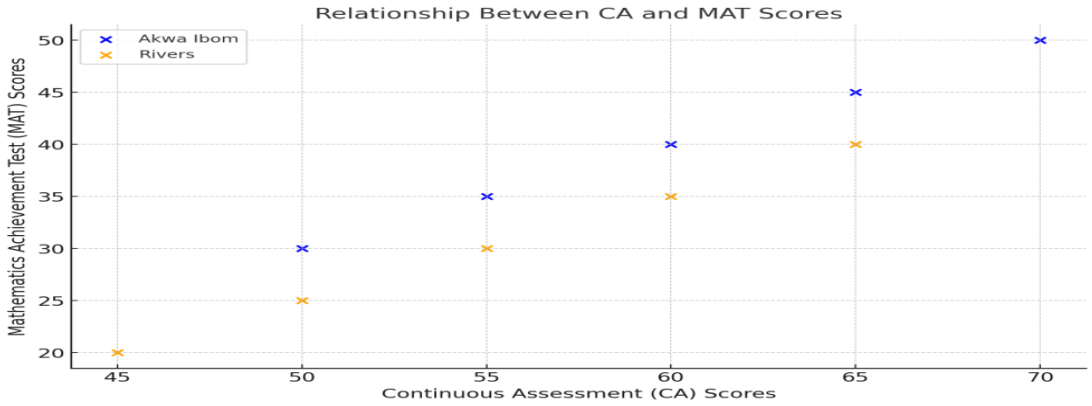


**Figure 3:** Scatter Plot showing Relationship Between CA and MAT Scores.

This scatter plot demonstrates the moderate positive correlation observed in both states: Akwa Ibom ($r = 0.42$) and Rivers ($r = 0.43$), both significant at $p < 0.05$. Each point represents a student's CA score (x-axis) and their corresponding MAT score (y-axis). This indicates that students with higher CA scores tend to perform better on the MAT, although the relationship is not perfect. The scatter plot visually demonstrates the trend and helps readers understand how CA scores moderately predict MAT performance. The plot supports the rejection of $H_{03}$, as a significant relationship ($p < 0.05$) exists between CA and MAT scores.

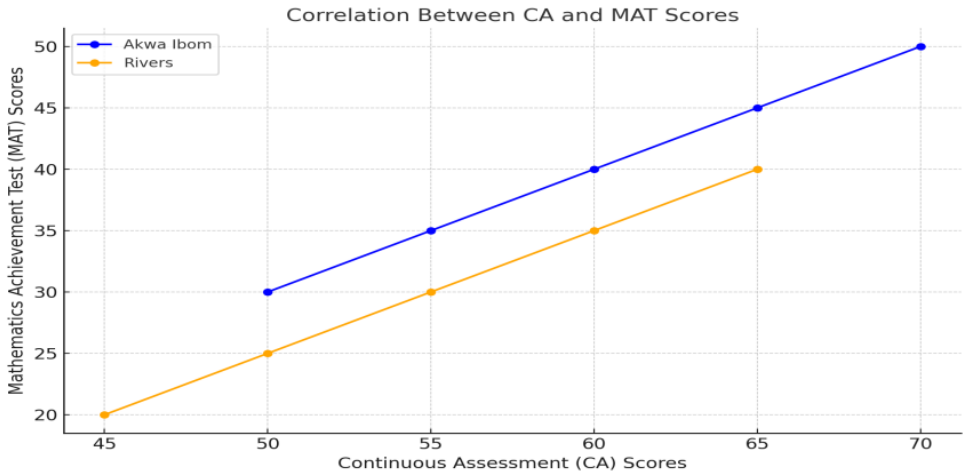


**Figure 4:** Line Graph showing Relationship Between CA and MAT Scores

The line graph in Figure 2 reveals a positive correlation between CA and MAT scores, indicating that higher performance in CA aligns with better achievement in MAT for both states. However, the strength of this relationship varies slightly, as reflected by MCA scores and MAT scores, state One was 0.42 and state Two 0.43 with associated probability of 0.00 for both. Histogram was included as additional visuals for broader comprehension. This visual corresponds to Research Question 3 and Hypothesis H03: Research Question 3: It illustrates the relationship between students' Continuous Assessment (CA) scores and their Mathematics Achievement Test (MAT) scores.

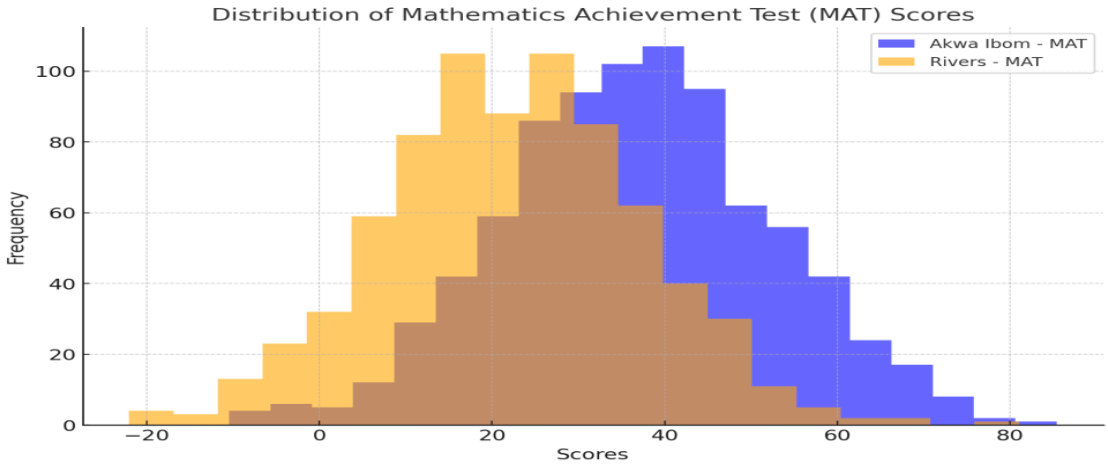


**Figure 5:** Histogram of CA Scores for both States

The histogram above helps show the distribution and variability of CA Scores in Akwa Ibom and Rivers, adding context to the mean differences observed.

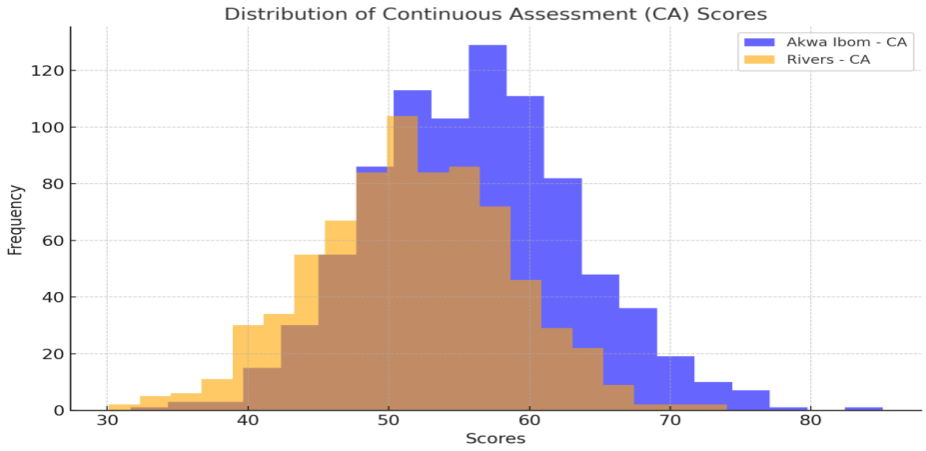


**Figure 6:** Histogram of MAT Scores for both States

Histogram of MAT Scores for both states displays the distribution of MAT scores, complementing the scatter plot by visualizing the performance spread.

The analysis of the robustness of test score equating for standard maintenance and comparison of continuous assessment scores in South-South Nigeria shows that, the mean ability estimates of students in state One and Two did not showed no significant difference was found in ability estimates between states (mean difference = 0.03; $t = 1.22$, $p > 0.05$), and that, the difference in the mean ability estimates and other moments (standard deviation and variance) did not show any score difference that matter for concurrent calibration but showed for linear calibration (SDTM). The result of t-test statistic indicated that the mean ability estimated from calibrated test form A1 and B1 using concurrent were statistically equivalent and was statistically not equivalent for linear equating. A significant mean difference (4.42, $p < 0.05$) favored Akwa Ibom, indicating higher ability estimates compared to Rivers. Relationship between MCA and MAT indicated that there exists a moderate positive correlation ($r = 0.42–0.43$, $p < 0.05$) between MCA score and MAT score. The r-value for state one was 0.42 with associated probability of 0.00 and state 2 was 0.43 with associated probability of 0.00. The associated probability in each case was less than 0.05 level of significance with 853 and 750 degrees of freedom respectively.

The graph visually compares ability estimates derived from concurrent calibration and linear equating for both states.Observing any deviations or overlaps in the lines helps identify if the equating methods lead to significant differences in mean ability estimates.For example: The lines for Akwa Ibom and Rivers overlap significantly in concurrent calibration but diverge in linear equating, it suggests a potential methodological impact on maintaining score equivalence. It supports the Results section by: Illustrating whether the methods are robust across the states. Highlighting whether one method (e.g., linear equating) introduces disparities. The scatterplot depicts the strength and direction of the relationship between CA scores and MAT scores for the two states. The visual representation helps test this hypothesis by showing whether a positive, negative, or no correlation exists. The scatterplot visually demonstrates the correlation between CA and MAT scores, indicating whether they are positively related: A positive linear trend suggests that students performing well in CA tests tend to also perform well in MAT. The distinct lines for Akwa Ibom and Rivers allow comparison between the states, revealing any discrepancies in the strength of the relationship.

This implies there is a significant relationship between the achievement of students in mathematics continuous assessment (MCA) test and mathematics achievement test (MAT) and the null hypothesis was not accepted. Concurrent calibration ensures score equivalence across test forms but may not detect subtle regional differences.Linear equating highlights disparities, emphasizing its suitability for comparing CA scores across regions. This outcome supports the works of Von Davies and Kong (2005); Arai and Mayekawa, 2011; Ayodele and Balogun, 2017; Ofqual, 2010; Agah, 2013; Kolen and Brennan (2004) who found out that, the mean difference between the two sample groups indicated a fairly large difference

between the two-population used in their study. Their studies also confirm the view that, the accuracy of the comparability of standards of the different forms of the same test over time or between test sessions depends to a certain degree on the accuracy and stability of item parameters, which are affected by a range of factors like, sample characteristics and size, common items, violation of model assumptions, model fit, curriculum change and population change, calibration methods and equating methods. Also as indicated earlier, one of the possibilities to address the difficulty in maintaining standards in continuous assessment testing is the use of IRT and CTT in linking and equating different test forms.

## V. CONCLUSION

The study aimed at ascertaining the robustness or appropriateness of three test scores equating methods (Linear equating, separate calibration and concurrent calibration) in the standard maintenance and comparison of students' continuous assessment measures. The research review shows that several methods are available for test equating and each method is unique for a given purpose. Using test equating as a means for comparison of students' performance requires standardized test, special design and software for data analysis. In view of the nature of CA testing, the wide application of IRT and CTT in educational measurement, particularly in test equating, and the availability of complex IRT analysis software, coupled with increasing IRT expertise in awarding organizations, the use of IRT and CTT could prove to be one of the viable approaches to maintaining standards over time or between test sessions in a CA measures.Robust equating methods, particularly linear equating, offer a practical solution for standardizing CA scores, ensuring fairness and accuracy in student performance evaluations. The study advances knowledge on applying equating methods to Nigerian CA practices, highlighting linear equating as a viable solution for standard maintenance. While linear equating effectively addresses disparities, advanced methods like equipercentile equating could provide a more comprehensive evaluation by accounting for nonlinear relationships between scores.Hybrid approaches that combine linear and equipercentile methods could further enhance score comparability.Future studies should explore these techniques to determine their feasibility and effectiveness in the Nigerian educational context. Other Recommendations are, adoption of Robust Equating Methods: Educational policymakers should adopt test equating methods like linear equating to standardize CA scores. Capacity Building for Teachers: Organize training programs to equip teachers with the knowledge andskills to implement equating methods effectively, including standardized test design and equating software. Standardization of CA Practices: Develop national guidelines to ensure uniformity in CA practices and reduce teacher biases or score inflation. Expand Research Scope: Include other states and geopolitical zones in future studies to validate findings and enhance their generalizability.

## REFERENCES

[1] Afemikhe, O. A. (2007). "Assessment and educational standard improvement: Reflections from Nigeria". A paper presented at the 33rd Annual conference of the International Association for Educational Assessment held at Baku, Azerbaijan. September 16th – 21st 2007.
[2] Agah, J. J. (2013). Relative efficiency of test scores equating methods in the comparison of students' continuous assessment measures. A published Doctoral Thesis University of Nigeria, Nsukka
[3] Altonji, J. G. (2009) Constructing AFQT Scores that are Comparable Across the NLSY79 and NSLY97. Retrieved November 10, 2009 from http://www.econ.yale.edu/-F188/AFQTmatch.pdf
[4] Arai, S. and Mayekawa, S. (2011). A Comparison of quating method and linking designes for developing an item pool under item response theory. Behaviormetrika 38(1)1-16. 2011.
[5] Ayodele, C.S. and Balogun, E.A. (2017). Causal effect of process variables on mathematics continuous assessment practice in Ekiti – State. *International Journal of Interdisciplinary Research Methods* Vol.4, No.3, pp.1-10, October 2017 (www.eajournals.org) Print ISSN: ISSN 2398-712X, Online ISSN: ISSN 2398-7138
[6] Chong, H. Y. & Sharon E. O. P. (2005) Test Equating by Common Items and Common Subjects: Concepts and Applications. Practical Assessment, Research & Evaluation 10 (4), 1-19

[7] Emaikwu, S. O. (2014) Issues in the assessment of effective classroom learning in Nigeria.*Journal of Humanities and Social Science (IOSR-JHSS) 19(10) 1-8 e-ISSN: 2279-0837, P-ISSN:2279-0845.*
[8] Faleye, B.A., & Adefisoye, B.T. (2016).Continuous assessment practices of secondary school teachers in Osun State, Nigeria.Journal of Psychology and Behavioral Science, 4(1), 44–55. Practices. New York: Springer-Verlag.
[9] Makiney, J. D., Rosen, C., Davis, B.W., Tinios, K. & Young, P. (2003). Examining the measurement equivalence of paper and computerized job analyses scales. Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.
[10] Moulton, M. H. (2004). Weighting and calibration: Merging Basch Reading and Maths Subscale Measures into a composite measure. Retrieved May 12, 2019 from www.aobfoundation.org.
[11] National Teachers' Institutes (2015). Manual for re-training of primary school teachers on school-based assessment. NTI press, Kaduna-Nigeria.
[12] Nwoke, B. I., Osuji, C. U., & Agi, U. K. (2017). Influence of computer-based test (CBT) on examination malpractice in public examinations. IOSR Journal of Research & Method in Education (IOSR-JRME), 7 (2Ver.2), 80-84.
[13] Ogunleye, A. W. and Omolayo, O. V. (2016). Classroom assessment in secondary schools in Nigeria. *International Journal of Social Science 5*(1), 1–6.
[14] Raju, N. S., Laffitte, L. J., & Byrne, B.M. (2002). Measurement equivalence: A comparison of confirmatory factor analysis and item response theory. Journal of Applied Psychology, 87(3), 517-529.
[15] Schumacker, R. E. (2005). Test Equating. Retrieved March 21st 2009 from
a. http://www.appliedmeasurementassociates.com/white%20papers/TEST%EQUATI NG.Pdf.
[16] The Office of Qualifications and Examinations Regulation [ Ofqual] (2010). Maintaining standards in on-demand testing using item response theory. The Qualifications and Curriculum Authority (QCA), an exempt charity under Schedule 2 of the Charities Act 1993. Spring Place Coventry Business Park Herald Avenue Coventry CV5 6UB.
[17] Uko, M. P. (2021). Wrıtıng To Learn and Trıangulatıon Assessment Approach on Academic Achievement, Self-Concept and Interest of Students in Mathematıcs in Colleges of Educatıon South-South Zone, Nıgerıa. An Unpublished Ph.D Dissertation Submitted To The Postgraduate School, Michael Okpara University Of Agriculture Umudike, Abia State, Nigeria.
[18] Van der Linden W. J. (2006). Equating scores from Adaptive to Linear Tests. Applied Psychological Measurement, 30(6), 493-508.
[19] Von Davier, A. A. & Kong, N. (2005). A unified Approach to Linear Equating for the Nonequivalent groups Designs. Journal of Educational and Behavioural Statistics, 30: 313-342.