Onuh, O., Obinne, A. D. E., Adulojo, M. O., & Emaikwu, S. O. (2024). Assessing Population Invariance of 2021 Agricultural Science Examination of National Examination Council (NECO) in Nigeria. *Journal of Research in Science and Mathematics Education (J-RSME), 3*(2), 64-75.

Original scientific paper https://doi.org/10.56855/jrsme.v3i2.980

Received: 21 March 2024.
Revised: 02 May 2024.
Accepted: 21 July 2024.

Assessing Population Invariance of 2021 Agricultural Science Examination of National Examination Council (NECO) in Nigeria

Omale Onuh[1*], Obinne A. D. E[2], Adulojo M. O[3,] Emaikwu S.O[4]

[1,2,3,4] College of Agricultural and Science Education, Joseph Sarwuan Tarka University, Makurdi, Nigeria

Abstract

Purpose: This study focused on fostering critical thinking in physics using graphic organizer-enhanced and context-based learning strategies among secondary students in Taraba State, Nigeria. **Methodology:** Five objectives guided the study using a quasi-experimental design. The sample comprised 225 students (males=113, females= 112). Data were collected using the Test of Critical Thinking Skill Acquisition (TOCTSA). The Kendal Tau-b inter-rater formula was used to determine the reliability of TOCTSA. Mean, standard deviation, and ANCOVA were the statistical methods employed. **Findings:** The study revealed a significant difference in the mean critical thinking scores in physics for students taught using the graphic organizer-enhanced learning strategy compared to those taught using the conventional strategy (F1,172 = 174.230; p = 0.000 < 0.05), as well as for those taught using the context-based learning strategy compared to the conventional strategy (F1,169 = 7.772; p = 0.006 < 0.05). However, there was no significant difference in the mean score of critical thinking in physics between male and female students taught using the graphic organizer-enhanced strategy (F1, 50 = 2.897; p = 0.095 > 0.05). Conversely, there was a significant difference in the mean score of critical thinking in physics between male and female students taught using the context-based learning strategy (F1,47 = 17.578; p = 0.000 < 0.05). **Significance:** These strategies have the potential to enhance the academic achievement of both male and female students and promote the acquisition of critical thinking skills in physics.

Keywords: *non-equivalent anchor test, population invariance, school location, test administration*.

© 2024 by the authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

[*] Corresponding author: *Omale Onuh,* omaleonuh@gmail.com

Introduction

The essence of testing in the school system is to reveal the latent ability of the examinee and to make grounds for assessment across the country to be as fair as possible, which may be lacking in many measuring instruments. One of the primary purposes of tests in our educational system is to provide a means of measuring or evaluating a group of examinees' abilities and skills that is as fair and objective as possible. Test has been fully accepted in most modern societies as the most objective method of decision-making in schools, industries, and government establishments.

Test/Examination plays an important role in learning for both students and their teachers. They are used to measure students' knowledge, intelligence, or other characteristics systematically. There are many reasons why teachers give tests to students. Teachers give tests to discover the learning abilities of their students to see how well students have learned a particular subject; some tests help the students to choose a vocation, and other tests help them to understand their own personality (Azpsychology, in Agbir, 2021). Though test result is accepted to be used in most societies as one of the most objective methods of decision-making, the use of test has sparked some concerns among the members of the public in recent years. These concerns have eroded people's faith in the power and efficacy of tests. Most of the serious allegations levied against tests pivot around the social issue that tests may show variation among different populations meant to take the test. Also, Emaikwu (2012) asserted that some tests unfairly favour examinees of a particular group. A test must not vary among any segment of the population taking the test.

Since the goal of testing is to reveal the latent ability of examinees, this latent ability is determined from the number of correct answers made by an examine and reported as a raw test score or some norm-based transformation of it. When raw scores are standardized through any transformation, scaling, or equating process, the resulting scores are known as measures (Altonji, 2009). These measures are statistical procedures used in interpreting scores. The statistical characteristics of tests used for assessment vary across schools and depend on the characteristics of the population they were designed for. These characteristics also vary according to examination bodies, school location, year of Administration and School type.

School location refers to the particular place, concerning other areas in the physical environment (rural or urban), where the school is situated. In Nigeria, rural life is uniform, homogeneous, and less complex than that of urban centres, with cultural diversity, which often is suspected to affect students' academic achievement (Adewale, 2015). In the same vein, studies have shown a positive influence with regard to school location, while others have shown the negative influence of school location on the student's learning outcome or achievement. The finding of Adebule and Aborisade (2013) indicated that students that resided in urban centres especially where there is a higher institution like polytechnics or universities are more likely to pursue higher education than those in the rural setting. School type has also been identified as a major factor contributing to academic performance and achievement in the school system. Public and Private schools are institutions owned and managed by public and individuals just as the names denote. These may lead to the examination score favouring one group over the other group. It has been observed that male students perform better in Agricultural Science than female students whereas some studies have shown that students in rural area perform better than their counterparts in an urban area. Some people have the presumption that students in public schools should perform better than students in private school in Agricultural Science subject, this notion could lead to bias decision and faulty conclusion being taken on such students.

Given these variations, measurement errors are bound to occur from tests developed by individual examination bodies and the scores generated from them. Scores obtained from different tests or examinations are often added up by these bodies to get an examinee's total score. Sequel to that, Agah (2015) asserted that 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 done over the achievement of students using test scores, these scores should be standardized or transformed through an appropriate method that is fair and objective. When scores are standardized using appropriate methods, they can be compared among students, classes/levels, years, or sessions. The issue of comparability of standards, which entails uniformity and quality of assessment instruments used as well as honesty and integrity in reporting test scores results among others has been a problem in Nigeria's educational system (Ayodele, 2013; Agah, 2015). This is the reason behind the study.

Test scores are of paramount importance and have consequential effects on the student's future. In Nigeria, examination bodies involved in high-stake testing programmes whose test scores have consequential effects on the students' future include National Examination Council (NECO), West African Examination Council (WAEC), National Business and Technical Examination Board (NABTEB), and Joint Admission and Matriculation Board (JAMB). These examination bodies are saddled with the responsibility of testing students, often administer their tests sometimes more than once a year and for security reasons they cannot use the same test form over different administrations. This requires the production of several test forms to be used in one testing session/ year. WAEC, NECO, and NABTEB have two testing sessions in a year (Internal and External Examinations) whereas JAMB administers once, but in several forms (Parallel forms). The administered forms apart from satisfying security demands are also used to monitor educational growth or trend in students' academic achievement.

The problem that arises as a result of having multiple test forms according to Baghaei (2010) is the difficulty level of the test forms and the comparability of the abilities and skills of examinees who take different forms, populations of the student who take the examination, school type (private or public) location of the school. These variables are part of this present study.

Two test forms can be different in difficulty even if developed based on similar content, format, and range of difficult questions asked, but the actual questions used might all be different in each form, hence accounting for tests with differing difficulty levels (Uysal & Kilmen, 2016). It is on this basis that Metibemu and Jonathan (2018) opined that if different examinees take different test forms, comparison among them is not possible. In the same vein, Aye and Htet (2010) asserted that any comparison of the raw scores on the two forms of the test would be unfair to the examinee who took the difficult form as they may have comparable low reported scores and grades on the examination/test. Also, since the population may differ considerably, it is difficult for comparison to be made based on pass or failure.

Despite this, the examination bodies that administer these alternate forms (parallel) of an examination use the raw scores of examinees to inform the decision on passing grades, admission to universities, award of scholarships, certification and other purposes without comparing them on a common scale. This practice used to put many eligible candidates who could have gained scholarships, university admission, and improved grades at a serious disadvantage because of having to take a difficult test, as compared to those who took a relatively easy examination. However, it is crucial for assessment programmes that include various types of tests to ensure that these test forms are constructed to be comparable. Parallel or alternate forms are considered comparable when they are designed to have similar content, statistical characteristics, and exhibit reliability. (Asiret & Sunbul, 2016). Thus, if two examinees who take parallel examination forms have different reported scores based on the type of test taken, then the concern is that the forms differ somewhat in difficulty. In this regard, if there is to be any comparison among examinees, a process that could be fair and objective in producing comparable scores on test forms of differing difficulties must be sought. Hence, there is the need for this study which is aimed at investigating population invariance of senior secondary certificate examination in Agricultural Science of NECO in North Central Nigeria.

Under a very specific set of conditions (the tests measure the same construct, generate scores with equal reliability, and hold a similar linking regardless of the population used to conduct the linking) a linking can lead to completely interchangeable scores, in which case the linking is referred to as an equating (Von Davier, 2013). Hence the focus of this study seeks to equate both internal and external agricultural science of 2021. Equating is a process used in comparing test scores of more than one test forms administered to a group of examinees. It is an empirical procedure used in establishing the relationship between the raw scores of two or more test forms. It enables the scores from one form of a test to be expressed in terms of the scores from the other form (Aye & Htet, 2010). Kolen and Brennan (2014) also conceived of test score equating as a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably. Population invariance refers to the situation in which a single linking function produces comparable score across scales for all sub-populations of interest. This implies that a test should be fair to all the sub-populations taking the test, when the school or the group who take the test are equated on the same scale and have the same equating function. The population invariance requirement demands that an equating result be independent of the unique characteristics of the examinee samples (for example males/females) used in the equating process. Powers in Agbir (2021), and Wyse and Reckase (2011) posited that no matter which groups of examinees are used, the equating results should not change with the characteristics of the particular examinee groups. The condition of population invariance is one of the ultimate goals of test equating. Thus, if population invariance is not obtained, the test forms may not measure the same construct as in the case of NECO Agricultural Science in June/July and November/December 2021.

Research Questions

The following research questions were posed to guide the study:

1. How invariant are the equated scores of the 2021 NECO Agricultural Science examinations in North Central Nigeria?
2. How invariant are the equated scores of the 2021 NECO SSCE Agricultural Science Internal and External examinations based on location in North Central Nigeria?

Research Hypotheses

The following research hypotheses were framed to guide the study and would be tested at 5% level of significance:

1. There is no significant difference between the mean invariance scores of the 2021 NECO Agricultural Science students in internal and external examinations.

2. There is no significant difference in the mean Invariance scores of the 2021 NECO Agricultural Science students based on location.

Gübeş and Kelecioğlu (2017) investigated group invariance of equating results which the authors check to find out if group invariance of equating functions means that equating is the same for everyone in the population. An equating study was conducted using an equivalent group design. The study group consisted of 15270 and 15323 9th-grade students who had taken Booklet B and Booklet D. The data used in this study were national assessment that was used for assessing elementary and secondary grade students' achievements in Turkish, mathematics, science, social science, and English domains in Turkey. The researchers used data from the 9th grade ÖBBS, which was administered in 2009. To prevent copying and allow for the sampling wide range of content, four different booklets (A, B, C, and D) were used in ÖBBS. There are 15 questions in each test for a total of 75 questions in one booklet. While choosing data for this study, the researchers conducted a dimensionality assessment (principal component analysis). The results showed that the IRT true-score equating method was more group-sensitive than the IRT observed-score equating method with respect to subgroups of self-perceived competence in geography and history lessons. Under all subgroups, the IRT true-score equating hadlarger REMSD values than the IRT observed-score equating method.

Agah (2015) used three equating methods in a study to ascertain the relative efficiency of test score equating methods in comparing students' continuous assessment measures. Linear equating method based on CTT and separate and concurrent calibrations based on the IRT framework were studied. Eleven research questions and six hypotheses guided the study which made use of the non-equivalent anchor test (NEAT) group design in Cross-River (State A) and Rivers State (State B) Nigeria. The population for the study comprised all senior secondary three III students of the 2010/2011 academic session in both states from which a sample of 2, 905 students was drawn through a multi-stage sampling procedure. The instrument for data collection was a 40-item two parallel forms of Mathematics Achievement Test (MAT) with reliability coefficients of 0.83 and 0.89 respectively. Data collected were analyzed using descriptive statistics, IRT differential item functioning likelihood ratio (IRTDIFLR) for 3 PLM, independent t-test, chi-square statistic, BILOG-MG 3 computer software, Pearson Product Moment Correlation Coefficient as well as Average Root Mean Square Error Difference (ARMSED). Findings from the study showed that Linear equating yielded the least ARMSED value amongst the three equating methods and was therefore deemed the most efficient in the study, also a significant difference did not exist in the item parameter estimates of the two parallel forms of MAT used for the test equating.

LaFlair et al. (2017) carried a study on equating in small-scale language testing programs. The study compared seven (7) equating methods namely: Mean, Linear, Levine, Linear Tucker, Chain equipercentile, circle-arc, Nominal weights mean and synthetic method to no equating (identity equating) based on small sample sizes. Two research questions guided the study. A non-equivalent group anchor test (NEAT) design was used to compare two listening and reading test forms based on small sample sizes (one with 173 test takers and the other 88) at a university's English for academic program (EAP). A 30 and 35 multiple-choice item instrument for listening and reading skills respectively were scored to generate data for the study. Descriptive statistics and correlation statistics were used to answer the research questions, while the equating methods were evaluated using the Standard Error of Equating (SEE). The results from the study revealed that among the seven equating methods, the circle arc produced the least equating errors. The result from the study revealed that the ability estimates of test-takers did not vary across the two forms of tests under the mean and circle arc methods of test equating.

Atsua et al. (2018) in a study on equating the 2015 and 2016 Basic Education Certificate Examination (BECE), compared the scores of candidates of JSS III students who sat for BECE in the 2015 and 2016 sessions in Civil Education using classical test theory equating method in Ibadan North, Oyo state. The study used the linear equating method. Five research questions guided this study. A survey design using a single group equating method was adopted. The population of the study was all the JSS III students in Ibadan North from where a sample of 619 students participated in the study. A 60-item multiple-choice test instrument for 2015 and 2016 was used for data collection. The research questions were answered using Stout's test of essential dimensionality, tetrachoric correlation, Chi-square goodness of fit built-in BILOG-MG3, and IRTPRO software. The findings from the study showed that the test data obey the assumption of unidimensionality. The results also showed that when test items were modelled using 1-PL, 2-PL, and 3-PL, the smallest chi-square value was observed when the data set was modelled with a 3-PL model. Findings from this study also indicated that the ability estimates of students using the two tests did not differ. The findings from this study also revealed that students' ability estimates when their scores were equated using the Linear equating method did not differ significantly.

Oluseyi (2018) conducted a study, which equated the state-unified and West African School Certificate Mathematics Examination items. The study determines the item parameters of the 2015 Unified Mathematics Examination and the WAEC Mathematics Examination as well as determines the comparability of the two Mathematics Examination items in terms of examinees' scores and item parameters. The finding showed that the difficult index of the state unified examination Mathematics items ranged between 1.34% to 50.00%, discrimination index ranged between -0.001 to 0.624. On the other hand, the difficult index of WAEC mathematics items ranged from 29.9% to 64.4%, while the discrimination index ranged from 0.344 to 0.885. Moreover, the difference in the difficulty and discrimination indices of both examinations were significant ($t = 8.682$ $p < 0.05$) and ($t = 16.664$, $p < 0.05$) respectively. Findings also showed that students' mean performance in the two examinations using Linear equating was significant ($t = 4.664$, $P < 0.05$).

Agbir (2021) assesses the relative efficiency of test score-equating methods in comparison to students' achievement test scores in Benue State. The author used four equating methods namely: Mean, Linear, Separate, and Concurrent Calibrations based on classical test, and item response theory were evaluated. Nine research questions were posed and seven hypotheses were formulated to guide the study. A random equivalent group design was adopted for this study. The population of the study comprised all Senior Secondary Two students in the 2018/2019 academic session from which a sample of 2901 students was drawn through a multi-stage sampling procedure. The instruments used for data collection in this study were two parallel forms of Chemistry Achievement Tests with reliability of 0.86 and 0.85 respectively. Data collected for the study were analysed using Jmetrik, NOHARM, and SPSS packages. The research questions were answered using descriptive statistics, chi-square goodness of fit, and NOHARM. The seven hypotheses were tested at 0.05 level of significance using t-test and chi-square statistics. The major findings of the data analysed revealed among others that, the root mean square error (RMSE) values obtained for the mean equating method were the least amongst the equating methods of Linear, Separate, and Concurrent calibrations used in this study and hence deemed to be more efficient in this study, also that CATs A and B were all unidimensional parallel tests.

Adewale (2016) in a study on equating two-year Basic Education Certificate Examinations (BECE), compares the scores of candidates of JSS 3 students who sat for BECE in 2013 and 2014 sessions in Basic Science and Technology in Oyo State. The study used two equating methods namely: Linear and equipercentile methods. Four research questions guided the study which also adopted a descriptive survey as a design. All 217, 651 students in the state for the two sessions (2013=108, 690; 2014=108, 961) participated in the study. A 60–item multiple-choice test instrument for each of the years was used. The research questions were answered using mean, standard deviation, T-scores, percentile ranks, and coefficient of variation statistics. The results of the study showed that the mean performance of students in the two years and two examinations was not significant. Also equated scores for the candidates for the two years were found to be equivalent. The findings also showed that the Linear equating method is more robust than the equipercentile equating method because it has a lower coefficient of variation and, hence, can be used for equating different examination scores.

Li (2015) evaluated the impact of construct shift on item parameter invariance, test equating, and proficiency estimates. In this study, the potential effect of multidimensionality on item parameter invariance and IRT model fit at the item level was examined. Two administrations representing two consecutive years of a statewide assessment were simulated using the operational item parameters and ability parameters. The study used the non-equivalent-group anchor test (NEAT) design with sample of 5000 examinee ability parameters for the multidimensional model with an external anchor of 21 linking items. To be specific, two test forms were simulated, and linking items were assigned to each form. Each form of the test consisted of 42 scoring items that included 32 dichotomously scored multiple-choice items, 6 dichotomously scored short answer items, and 4 polytomously scored open response items. The equating items were composed of 16 multiple-choice items, 3 short answer items, and 2 open-response items. In the study, construct shift occurred in the operational scoring part in Form Y and not in what was expected to be the more consequential portion of the test—the linking items Scoring items were randomly selected to introduce multidimensionality that appeared as multiple-choice item type, measuring three different traits. Data generation was performed with the computer program *R*, and the finding highlighted the fact that very modest changes in construct shifts in the operational items of a test would not be problematic. It was found when a construct shift occurred in the operational items, the impact was minimal in most aspects. Whereas, when construct shift occurred in the anchor and when these equating items were included in the total test, the influence of construct shift was quite substantial. The study provided evidence of consequences caused by construct shift as a function of the correlations between pairs of dimensions, the amount of shift, and where the shift occurs (equating or operational items). In general, it was found that construct shift had an increasing impact as the amount of shift increased and the correlations between dimensions decreased.

Kelecioglu and Ozturk (2013) in a study compared Linear and equipercentile equating methods aimed at equating raw scores of 9th grade 2009 in the Turkish National Examination in Social Sciences using test forms A and C with Linear and equipercentile equating methods. The study which adopted a random group equating design used a sample of 16,670 and 15,743 9th grade students who took two parallel test forms A and C respectively. The reliability of the two parallel forms which consisted of 15 multiple-choice items each was established using $K\text{-}R_{20}$, and a value of 0.76 and 0.75 were obtained for test forms A and C respectively. RMSD was used in the study to compare random errors of equating the two methods. The result from this study indicated that there was a linear relationship between equivalent scores of form A and a raw score of forms C and form C was easier than form A along the score scale. Also, the linear equating method produced lower equating errors than the equipercentile method.

Method

The study adopted a non-equivalent anchor test (NEAT) group equating design. Since the tests are parallel but are administered at different intervals or periods, the two groups of students are not equivalent in terms of ability. The population of the study was 97,413 Senior Secondary students who wrote Agricultural Science in 2021 academic sessions in North Central. The students who sat for NECO SSCE 2021 internal were 89,680 while those who sat for SSCE 2021 external were 7733. The sample size for this study was 2682 Agricultural Science students with 1506 for internal 2021 and 1176 for external 2021 NECO examination. One instrument was used for data collection and it was a self-structured proforma titled ''NECO Agricultural Science OMR Score Retrieval Proforma. Data were analysed using the IRT Statistical tool and software. Research questions one and two were answered using coefficient of variation statistics. The hypotheses were tested using a t-test at 0.05 level of significance. The decision rule to reject or not to reject the hypotheses were based on the set value of 0.05, where the P-value is greater than the set value of 0.05(P>0.05) the hypothesis would not be rejected but was rejected when the P-value was less than the set value of 0.05(P<0.05).

Results and Discussion

Research questions 1: How invariant are the equated scores of the 2021 NECO Agricultural Science examinations in North Central Nigeria?

In order to answer this research question, respondents' equated scores in each of the examination were independently summed and means and standard deviations were computed in order to calculate co-efficient of variation of the examination. The normalized measure of dispersion of a probability distribution is called as coefficient of variation and abbreviated as CV. In probability theory and statistics, it is also known as unitized risk or the variation coefficient. The CV is derived from the ratio of the standard deviation to the non-zero mean and the absolute value is taken for the mean to ensure it is always positive. It is sometimes expressed as percentage; in which case the CV is multiplied by 100, i.e Coefficient of variation = Standard deviation/ Mean ×100.

Table 1. Invariance of equated scores of Agricultural Science of Internal and External 2021

	Mean	S.D.	Coefficient of Variation
2021 Internal	31.48	17.31	55
2021 External	29.08	17.43	60

Table 1 revealed the same mean values for the examinations with standard deviations and their coefficient of variation. The coefficient of variation of 55% and 60% were respectively observed. The lower the coefficient of variation, the more precise the estimate, on the other hand the higher the coefficient of variation the less precise the estimate. Thus, internal NECO is preferable because it has lower coefficient of variation. This implies that the two tests are not on the same scale and they are not comparable, that the score are at variance with each other.

Research Question 2: How invariant are the equated scores across the 2021 Agricultural Science NECO SSCE Examination based on Location in North Central Nigeria?

Table 2. Invariance of Equated scores of Agricultural Science examination of 2021 based on location

	Mean	S.D.	Coefficient of Variation
Urban	28.49	6.26	22
Rural	29.09	7.30	25

Table 2 shows the mean values of Agricultural Science examination based on location with standard deviations and their coefficient of variation of 22% and 25% were respectively observed for urban and rural schools. This implies that the scores are at variance with each other. The corresponding research hypotheses were tested using independent t-test at 0.05 level of significance and the results were presented accordingly.

Hypothesis One: There is no significant difference between the mean Invariance scores of the 2021 NECO Agricultural Science students in internal and external examinations.

Table 3. Independent t-test Analysis of mean Invariance of Score of Internal and External Student NECO Examination

Examination	n	Mean	Standard Deviation	Mean Difference	t-cal	α	df	P –value	Remark
Internal	1506	31.48	17.13	2.40	3.55	0.05	2680	0.00	S,R
External	1176	29.08	17.43						

P < 0.05; *Significant; Ho is rejected.*

Table 3 shows the independent t-test analysis of students' invariance based on NECO internal and external examinations of 2021. The analysis showed that, there was significant difference in the invariance equated score of internal (M = 31.48, SD = 17.13), and external (M = 29.08, SD = 17.43), t (3.55) =DF=2680, p = 0.00. The null hypothesis of no significant difference was rejected. This result means that, the invariance equated score based on the examination type in terms of internal and external examinations differ significantly.

Hypothesis two: There is no significant difference in the equated invariance scores of the 2021 NECO Agricultural Science student based on location.

Table 4. Independent t-test Analysis of equated Invariance scores of the 2021 NECO Agricultural Science students score based on location

Location	N	Mean	Std Deviation	Mean Difference	t-cal	α	df	P – value	remark
Urban	1341	28.49	6.26	-0.600	-2.285	0.05	2680	0.022	S,R
Rural	1341	29.09	7.30						

P < 0.05; *Significant; Ho rejected.*

Table 2 shows the independent t-test analysis of the equated invariance scores of the 2021 NECO

Agricultural Science students based on location. The analysis showed that there was a significant difference in the invariance equated scores in urban (M = 28.49, SD = 6.26), and rural (M = 29.09, SD = 7.30), with t-cal value of -2.28 and with df =2680, p = 0.022. The null hypothesis of no significant difference was rejected. This result means that the invariance equated scores based on the location of urban and rural differ significantly.

From the analysis of result, the coefficient of variation of 55% and 60% were respectively observed for internal and external NECO examinations respectively. The lower the coefficient of variation, the more precise the estimate, on the other hand the higher the coefficient of variation the less precise the estimate. Thus, internal NECO is preferable because it has lower coefficient of variation. This implies that the two tests are not on the same scale and they are not comparable, that the score are at variance with each other.

Based on the findings of the study, it was discovered that the internal NECO has a lower coefficient of variation than external NECO examination. This is also buttressed by the hypothesis that there is a statistical significant difference between the equated score of internal and external examinations. This implies that the two tests are not on the same scale and they are not comparable. The result negates the findings of LaFlair et al. (2017), Adewale (2016), Kelecioglu and Ozturk (2013), Agah (2015), Atsua et al. (2018). Agbir, (2021) revealed that the ability estimates of test-takers did not vary across the forms of tests when equated. While the study is in consonance with the studies of Oluseyi (2018) and Gübeş and Kelecioğlu (2017) who found that there is a significant difference in the equated score of the forms of test.

From the analysis of result also, the mean values of Agricultural Science examination based on location with standard deviations and their coefficient of variation of 22% and 25% were respectively observed for urban and rural schools. This implies that the scores are at variance with each other. The result of hypothesis indicated that there was a significant difference in the invariance equated scores in urban (M = 28.49, SD = 6.26), and rural (M = 29.09, SD = 7.30), with t-cal value of -2.28 and with df =2680, p = 0.022. This result may be that the abilities of the students differ, also the present result indicated significant invariance difference between students in urban and rural areas. It can therefore be asserted that the 2021 NECO Agricultural Science examinations of Internal and external are not equivalent, and that the difficulty of both examinations differs considerably. It must be recalled that when comparisons are to be made, they should be based on the same scale, and the scores to be compared must be placed on the same equating scale to avoid invariance.

Conclusion

Based on the findings, it was concluded that the 2021 NECO Agricultural Science of internal and external examinations are not equivalent and that the difficulty of both examinations equally differs. The examination is also biased based on school location. It must be concluded that when comparisons are to be made, they should be based on the same scale, and the scores to be compared must be placed on the same equating scale to avoid invariance. The importance of tests which include fairness and equity for evaluation was not valid and this calls for a better scrutiny. In conclusion, for any test to be valid for decision-making, the assessment tool for standardization must be considered an important factor by examination bodies.

Recommendations

Based on the findings of this study, the following recommendations were made.

1. Population invariance should be used for the comparison of two parallel examinations using mean equating method since it is more efficient
2. Test experts and developers should explore the use of the IRT approach to develop standardized test items that are meant for public examination.

References

Adebule, S.O., & Aborisade, O.J. (2013). Influence of Study Interest and School Location on the Attitude of Secondary School Students towards Mathematics in Ekiti State, Nigeria. *Greener Journal of Educational Research, 3*, 229-232.

Adewale, J. G. (2016). Equating two years of BECE Results in basic science and technology in Oyo State, Nigeria. International Centre for Educational Evaluation (ICEE) Institute of Education, University of Ibadan-Nigeria.

Agah, J. (2015). Relative efficiency of test scores equating methods in comparison of students' continuous assessment measures. A Ph.D. Thesis. Department of Science Education, University of Nigeria, Nsukkka. Nigeria.

Agbir, J. D. (2021). Relative efficiency of classical test and item response theories in test scores equating methods using standardized Chemistry achievement tests. *Unpublished Ph.D. thesis. Department of Educational Foundations and General Studies, Federal University of Agriculture Makurdi Benue State.*

Altonji, J. G. (2009). Constructing AFQT Scores that are Comparable Across the NLSY79 and NSLY97. Retrieved from: http://www.econ.yale.edu/F188/AFQTmatch.pdf

Asiret, S., & Sunbul, O. S. (2016). Investigating test equating methods in small samples through various factors. *Educational Sciences: Theory and Practice*, *16*(2), 647-668. https://psycnet.apa.org/doi/10.12738/estp.2016.2.2762

Atsua, T. G., Uzoeshi, V. I., & Oludi, P. (2018). Equating 2015 and 2016 basic education certificate examination on civic education using classical test theory and item response theory in Oyo State. Nigeria *Journal of Pristine, 14*(1) 2250-9593.

Aye, M. A., & Htet, L. O. (2010). An application of linear test equating method in scoring. *Yangon Institute of Education Research Journal, 2*(1), 1 -15.

Ayodele, C. S. (2013). Transformation of continuous assessment scores among schools in Nigeria. *European Scientific Journal, 8*(26), 171-180.

Azpsychology. (2010). *Importance of testing in psychology and education*. Retrieved from: http://www.a2zpsychology.com/articles/importance_of_testing_in_psychology.htm

Baghaei, P. (2010). Test score equating and fairness in language assessment. *Journal of English and Literary Studies, 1*(3), 113 -128.

Emaikwu, S. O. (2012). Fundamentals of test, measurement, and evaluation with psychometric theories (2nd Edition). SAP Ltd (selfers Academic Press).

Gübeş, N. O., & Kalecioğlu, H. (2017). Investigating group invariance of equating results. *İlköğretim Online*, *16*(1), https://doi.org/10.17051/io.2017.34481

Kelecioglu, H., & Ozturk, N. G. (2013). Comparing linear equating and equipercentile equating methods using random group design. *International online journal of Educational Services, 5*(1), 227-241.

Kolen, M. J., & Brennan, R. (2014). Test equating: Methods and practices (3rd. edition). Springer-verlag.

LaFlair, G. T, May, L. D., & Arvizu, N. G. (2017). Equating in small-scale language testing-Programmes. *Language Testing 34*(1), 127-144. https://doi.org/10.1177/0265532215620825

Li, X. (2015). Evaluating the impact of construct shift on item parameter invariance, test equating and proficiency estimates.

Doctoral Dissertations. 497. Retrived from: https://scholarworks.umass.edu/dissertations_2/497

Metibemu, M. A., & Jonathan U. (2018). Evaluating the appropriateness of using tetrachoric correlation coefficient among items in the assessment of item local independence assumption of multiple-choice test. *Nigerian Journal of Educational Research and Evaluation*, *17*(1), 133-144.

Oluseyi, A. E. (2018). Equating the state unified and West African School Certificate Mathematics examination items. *American International Journal of Contemporary Research*, 8*(4)*, 100-106. http://dx.doi.org/10.30845/aijcr.v8n4p10

Uysal, I., & Kilmen, K (2016). Comparison of item response theory test equating methods for mixed format tests. *International online journal of Education Sciences, 8*(2), 1-11. http://dx.doi.org/10.15345/iojes.2016.02.001

Von Davier, A. A. (2013). Observed-Score Equating: An overview. *Psychometrika*, *78*(4), 605–623. https://doi.org/10.1007/s11336-013-9319-3

Wyse, A. E. & Reckase, M. D. (2011). A graphical approach to evaluating equating using test characteristic curves. *Applied Psychological Measurement, 35*(3), 217-234. https://doi.org/10.1177/0146621610377082

Zieky, M. (2006). Fairness reviews in assessment. In S. M. Downing, & T. M. Haladyna (Eds.), Handbook of Test Development (pp. 359-376). Mahwah, NJ: Lawrence Erlbaum Associates.