Comparison of African Indigenous and Western Intelligence Tests using Validation Processes of Bakare Progressive Matrix And Wechsler Adult Intelligence Tests

***Taiwo, Aebukola Kabir and Ojuolape, Mumud Olabode**
Department of Counselling and Human Development Studies, Faculty of Education, University of Ibadan
***Corresponding Author**

DOI: https://dx.doi.org/10.47772/IJRISS.2023.7657
Received: 21 May 2023; Accepted: 30 May 2023; Published: 06 July 2023

ABSTRACT

The study was designed to examine test bias: a comparison of indigenous and Western standardized intelligence test validation processes using the Bakare Progressive Matrix and Wechsler intelligence tests among adults in the Ibadan metropolis. The study was anchored on Van de Vijver (1998)'s Bias and Equivalence theory, while the survey design was adopted. A sample size of 550secondary school teachers participated in the study. The multi-stage sampling procedure was adopted. In the first stage, all five Local Government Areas (LGAs) in the metropolis, were enumerated. In the second stage, five schools were randomly selected from each of the LGAs. In the third stage, 22 teachers were randomly selected in each school, totaling 550 teachers. The instruments used were: Bakare Progressive Matrix (KR = 0.96) and Wechsler Adult Intelligence Scale IV (WASI-IV – KR = 0.93). Item parameters estimate and item analysis was used to analyse the findings. The internal consistency of each of the intelligence tests varies as the foreign test (KR21= 0.713) displayed a better reliability coefficient than the local test using the Kuder Richardson reliability index. The indigenous intelligence scale difficulty ranges between 0.30-0.965 (3.0%-96.5%), while the Western intelligence scale ranges between 0.000-0.99 (0.0%-99%). This implies that there is a need for improvement on the indigenous scale. It was recommended that efforts should be made by the government, research institutes, and educational bodies to ensure that the indigenous standardized scales are promoted and subjected to rigorous and sophisticated psychological testing that will promote replicative practices outside the shores of Africa that would be readily relevant, accepted and practiced in Western communities.

Keywords: Indigenous Standardized Test, Test Bias, Bakare Progressive Matrix, Wechsler Adult intelligence

INTRODUCTION

The urge to embark on this study was stimulated by the researchers' desire to find out the test bias of African indigenous and Western standardized intelligence tests. The issue of test bias has brought about differences in intelligence between African and other Western countries. This triggered more insightful questions such as; how recognized is the indigenous standardized test relevant to the Western standardized test, and why are indigenous tests not replicated outside the shores of Africa? Does indigenous standardized test predict over or alongside the Western standardized test? All these questions pose a significant problem to this study placing it on a controversial subject matter of test bias.

The main purpose of this study is to examine test bias: a comparison of indigenous and Western standardized intelligence test validation processes using the Bakare Progressive Matrix and Weschler Standardized Intelligence Test. The cultural bias in standardized tests is recognized as a problem to be avoided. The study intended in generating workable and operational items for the measures of intelligence. In line with the interest of this study, the researchers try to place this work as a limitation on the part of African indigenous standardized tests, although most Western scholars have not been able to promote African indigenous standardized tests on the contrary, African researchers have promoted Western standardized tests in so many studies and situations. Differences in intelligence must be understood culturally. It has already been shown that most IQ tests focus predominantly on the componential aspect of intelligence and neglect the other aspects of intelligence. This points to the fact that different cultures emphasize different types of intelligence and one should be culturally sensitive when assessing intelligence.

This research work will be based on Reynolds' (2000) assumptions which state that the most common explanations of racial and ethnic differences in intelligence fall into four categories: (a) the differences have a primarily genetic basis; (b) the differences have an environmental basis; (c) the differences are due to the interactive effect of genes and environment; and (d) the tests are faulty in such a way that minorities' true knowledge, skills, or aptitudes are systematically underestimated. Although research tends strongly to support the cross-cultural equivalence of intelligence tests, psychologists and scientists have not yet exhausted the research needed to be done to consider the cultural test bias hypothesis and its alternatives (Reynolds, 2000; Khan, 2019).

The research problem or gap emanated from previous studies was that although science strove for objectivity, the individual scientists were affected by their attitudes, which were, for the most part, a reflection of the culture or society in which they lived and worked. Not only were there different theories on intelligence, but there were also myriad differences in mean levels of the performance of non-whites on standardized tests of intelligence, aptitude, or academic achievement (Reynolds, 2000). In fact, as Reynolds (2000) and Brown, (2019)posit, "There are few issues as volatile and polemic in psychology as the issue of race and ethnic differences in mental test scores."

However, this research problem ignited a spirited debate over the genesis of racial and ethnic differences in performance on mental tests. The most valid explanation of racial and ethnic group differences, however, is the cultural bias hypothesis, which states that "group differences in mental test scores occur due to systematic underestimation of minority groups' aptitude levels; or, more generally, tests contain a systematic error that occurs as a function of what should be irrelevant nominal group membership (e.g., race, ethnicity, gender, and socio economic status Reynolds, 2000). So, it could be explained clearly that the issue of overestimation in the comparison of African and Western bits of intelligence lies in test biases (Braaten and Dennis, 2020).

Test bias has been occurring for years, there is a bias between the Indigenous test and Western intelligence test and this research work has been able to unveil the biases on the measure of intelligence tests using the Bakare Progressive Matrix test and Wechsler intelligence test. A test is said to be biased when it yields clear and systematic differences among the results of the **test**-takers.

It is a belief that the Western standardized test is more validated than the indigenous test and this has brought about a lot of poor interpretation of intelligence measures. Most of the research on intelligence measures in Nigeria has had its origin in Western research, thereby rendering Nigeria's local research to have its toll on foreign research, a situation that has introduced bias in our local findings and interpretations.

Indigenous assessments should be based on cultural experiences. On the other hand, efforts to indigenize assessments are geared towards making the so-called foreign assessments suitable for use in African settings. One may see indigenization as making the assessments appropriate taking into consideration the characteristics of those assessed while indigenous assessments are those having a theoretical foundation based on the local conditions. It has been observed that poor performance in public examinations can be attributed to not only the test takers but also the biases of the test (Whitaker, 2020).

Psychological assessment is pervasive in life and particularly in clinical and educational discourse. In the educational sphere, it assumes a central stage as most decisions that have to be made depend on the availability of current, valid, and dependable information gotten from the standardized test instrument. The decisions focus on issues related to the students, school system, school programmes, educational policy, psychological concepts of individuals, and the society at large as rightly observed by Nitko (2013) and Stough, Kerkin, Bates, and Mangan, (2020). Within the informal and non-formal system in Nigeria, psychological assessment also plays an important role but this is usually ignored ignorantly in discussions. Nonetheless, efforts to get the best out of education must be given the needed impetus on psychological assessment. Without it, one may not be sure whether the expectations and goals enshrined in educational programmes are being met. It is as a result that assessment has continued to be emphasized in education.

Tests are considered biased if a test design, or the way results are interpreted and used, systematically favour certain groups of students over others, such as students of color, lower-income backgrounds, those who are not proficient in the English language, or those who are not fluent in certain cultural customs and traditions. Identifying test bias requires that test developers and educators determine why one group of students tends to do better or worse than another group on a particular test. For example, is it because of the characteristics of the group members, the environment in which they are tested, or the characteristics of the test design and questions? As student populations in public schools become more diverse, and tests assume more central roles in determining individual success or access to opportunities, the question of bias and how to eliminate it has grown in importance.

Differences in intelligence, or rather, in intelligence types, must be understood culturally as well as biologically. It has already been shown that most IQ tests focus predominantly on the componential aspect of intelligence and neglect the other aspects of intelligence (Aaron, Dasgupta, and Kushan, 2020). This points to the fact that different cultures emphasize different types of intelligence and one should be "culturally sensitive" when evaluating intelligence. Intelligence also has a strong information-processing component, it may involve generating new ideas and it operates within a specific cultural context. Westerners believe that an averagely intelligent white manis equivalent to the most highly intelligent man in Africa, an assessment that has introduced biases.

The issue of indigenous standardized assessment tests has long been a very contradicting and controversial context of assessment in the field of psychology throughout the globe (Taylor, 2011). Previous research(Mitchell, 2018; Whitaker, 2020)conducted by scholars in the field of measurement and evaluation have tried to investigate the bigotry instincts of the general empirical stance in various areas of psychological attributes like intelligence, especially those channeled to the cognitive domain and affective domain. Increasingly, the issues of indigenous or indigenized assessment scales have been a major concern to stakeholders in the assessment and evaluation. The point of the comparison between the indigenous and Western standardized tests has not only cut across the various endeavor – cultural, psychological, educational, and societal acceptance but also, the issue of discrimination has been widely discussed. In this case, scholars across the globe have sorted for measures to reduce the tremendous crisis and concerns raised by replicating Western standardized tests to meet cultural demands, especially in areas of test assessment and measurement in psychology. Despite these researches, there is still a paucity of literature on African indigenous standardized instruments for measuring intelligence. The controversy over indigenous standardized tests lingers on while Western standardized tests are subjected to cultural unfairness and bias among Africans. This study, therefore, was designed to compare African indigenous and Western intelligence tests using the validation process of the Bakare progressive matrix and Wechsler adult intelligence tests.

This study was based on Van de Vijver (1998)'s proposition of Bias and Equivalence theory, which postulated that a measuring instrument is biased if its scoring is incongruent with its psychological meaning across the cultural groups in the comparison. Van de Vijver (1998) sighted, in his example, that individual differences in intelligence test scores may reflect differences in intelligence in a single cultural group, whereas intergroup differences may be largely due to differences in education and test experience.

Van de Vijver (1998) defined equivalence as the question of whether there is any difference in measurement level of within- and between-group comparisons. He postulated that, If the measure is biased against some cultural group, individual differences within a cultural population and across cultural populations are not measured at the same scale. He then sighted three characteristics that can be derived from these definitions as bias, which refers to unintended sources of variation that constitute alternative explanations of intergroup differences. He posited that If bias is present, cross-cultural score differences are not engendered by the target construct (e.g., intelligence or political affiliation) but by some other characteristic (e.g., social desirability or education). He affirmed that bias and equivalence are not intrinsic to an instrument but characteristics of a specific cross-cultural comparison. Van de Vijver (1998) stated that both instrument and sample characteristics will influence the likelihood of the occurrence of bias. In this context, the researchers in this study agreed with Van de Vijer (1998)'s proposition that an instrument used to measure intelligence in Western countries may be biased in a comparison of African countries. Van de Vijer (1998) made explanations that bias will often increase with the cultural distance to be bridged by the instrument and is also more likely when an instrument shows more cultural saturation.

METHODS

This chapter deals with the method to be used and describes the method to be adopted in carrying out the study. It describes the research design, population, sampling procedure, research instrument, validation procedure for the administration of the questionnaire, scanning of the instrument, and method of data analysis.

Design

The study adopts a cross-sectional and expo facto method to investigate test bias: a comparison of indigenous and Western standardized intelligence test validation processes using the Bakare Progressive Matrix and Wechsler intelligence tests among adults in Ibadan because it is considered one of the best available designs for purposes of describing a fairly large population. The research design is adopted because the researcher did not manipulate the variables of interest in the study.

Population

The population for this study constituted male and female adults in the Ibadan metropolis, Oyo state. The purpose of selecting this population was to examine the applicability of the test comparison of this study to the population of teachers with common characteristics. The Ibadan city is considered the most populous city of Oyo State, Nigeria. It is the third-largest city by population of people in Nigeria after Lagos and Kano, with over 6 million people within its metropolitan areas. Ibadan is ranked the second fastest-growing city on the African Continent (UN Human Settlements Research Program, 2022)

Sample and Sampling Techniques

The multi-stage sampling procedure was adopted. In the first stage, all five Local Government Areas (LGAs– Ibadan North, Ibadan North-east, Ibadan North-west, Ibadan South-east, and Ibadan South-west) in the metropolis, were enumerated. In the second stage, five schools were randomly selected from each of the

LGAs. In the third stage, 22 teachers were randomly selected in each school, totaling 550 teachers.

Instruments

The study made use of two validated scales to gather information from the respondents of the study. The scales were Bakare Progressive Matrix and Wechsler Adult Intelligence Scale IV (WASI-IV).

Bakare Adult Intelligence Scale: A scale developed by Charles Bakare, was employed to determine intelligence based on the progressive matrix of the participants. The scale has five (5) sections: A to E.

Wechsler Adult Intelligence Scale Fourth Edition (WAIS-IV): developed by David Wechsler (2008), was employed to determine the intelligence attributes of people based on interview responses to items showing the appraisal and expression of intelligence attribute resident in self. This instrument has been examined as suitable for senior secondary school students in Nigeria. The scale has fifteen (15) subtest tags: Block design, Similarity, Digit Span, Matrix Reasoning, Vocabulary, Arithmetic, Symbol Search, Visual Puzzle, Information Codding Letter-Number seq., Figure Weight, Comprehension, and Picture Completion but only the block design was used for this research work. The subtest was then categorized into four phases such as process score, composite scale, and full scale.

Validity and Reliability of the Instrument

The Wechsler Adult Intelligence is a well-established scale and it has fairly high consistency. Over a two to twelve-week time period, the test-retest reliabilities ranged from 0.70 (7 subscales) to 0.90 (2 subscales). Inter-scorer coefficients were very high, all being above 0.90. According to the test manual, the instrument targets psycho-educational disability, neuro-psychiatric and organic dysfunction, and giftedness. The WAIS correlated highly with the Stanford-Binet IV test (0.88) and had high concordance with various measures: memory, language, dexterity, motor speed, attention, and cognitive ability.

Bakare Progressive Matrix: This test is broad-based and consists of sixty-one (61) items. Construct validity was developed for the test, and this was done by establishing face validity, content validity, criterion-related validity, convergence validity, discriminant validity, internal consistency reliability, and factor analysis. Test broad-based yielded a high-reliability coefficient alpha of 0.92.

RESULTS

Research Question 1: What difference exists in the descriptive characteristics of foreign and local scales?

Table 1: Kelly Score summary showing Descriptive Statistics of the local and foreign intelligence test scores.

Kelly Score Descriptive Statistics			
Test 1: **Bakare Progressive test (Local)**		Test 2: **Wechsler Intelligence test (Foreign)**	
Statistic	Value	Statistic	Value
N	200.0000	N	200.0000
Min	1.3900	Min	2.7100
Max	5.8800	Max	9.9200
Mean	2.3521	Mean	9.7112
St. Dev.	0.6051	St. Dev.	0.6404
Skewness	1.5290	Skewness	-7.1278
Kurtosis	5.8229	Kurtosis	71.8918

KR21	0.540	KR21	0.713
Cronbach alpha	0.598	Cronbach alpha	0.720

Table 1 reveals that 200 testees participated in the study, the internal consistency of each of the intelligent tests varies as the foreign test (KR21= 0.713) displayed a better reliability coefficient than the local test using the Kuder Richardson reliability index. However, the local intelligent test reliability coefficient improved despite its weakness after the removal of item b12. The degree of skewness recorded by the two tests differs, the local test is mildly positively skewed while the foreign test is negatively skewed. By implication, the foreign scale appears too easy although with a kurtosis of about 72 score which shows many candidates got the test right.

The mean score recorded after test scaling revealed that from the foreign intelligent test (mean= 0.971), 97.1% of the testees got the test right while on the local test (mean= 0.235), 23.5% of the testees got the test right. By implication, the foreign intelligent test is easier for Nigerian testees than the locally-made intelligence test.

Research question 2: Do variances exist in the item parameter estimation of local (Bakare Progressive test) and foreign (Wechsler Intelligence test) intelligence tests?

To answer the question 14/10 items on intelligent tests were subjected to analysis using JMETRIK. As a result of the model fit assessment conducted, the 2-PL model represents the CCT statistics, **p** represents the item difficulty indices and $\mathbf{r_{pbs}}$ represents the discrimination indices (using the point biserial correlation).

Table 2: Item Parameter Estimate Summary showing difficulty index (*p*), discrimination indices (r_{pbs}) of Bakare Progressive test and Wechsler Intelligence test

Test 1: Bakare Progressive test					**Test 2: Wechsler Intelligence test**				
Item	Option (Score)	**p**	Std. Dev.	$\mathbf{r_{pbs}}$.	Item	Option (Score)	**p**	Std. Dev.	$\mathbf{r_{pbs}}$
b1	Overall	0.4150	0.4940	0.2322	c5	Overall	**0.9400**	0.2381	0.4540
	Wrong(0.0)	0.5800	0.4948	**-0.7055**		0(0.0)	**0.0550**	0.2286	**-0.6137**
	Right(1.0)	0.4150	0.4940	0.2322		1(1.0)	**0.9400**	0.2381	0.4540
b2	Overall	0.5850	0.4940	0.1577	c6	Overall	**0.9950**	0.0707	0.7417
	0(0.0)	0.4100	0.4931	**-0.6648**		0(0.0)	**0.0000**	0.0000	NaN
	1(1.0)	0.5850	0.4940	0.1577		1(1.0)	**0.9950**	0.0707	0.7417
b3	Overall	**0.1400**	0.3479	0.4107	c7	Overall	**0.9250**	0.2641	0.4006
	0(0.0)	**0.8550**	0.3530	**-0.6975**		0(0.0)	**0.0700**	0.2558	**-0.6302**
	1(1.0)	**0.1400**	0.3479	0.4107		1(1.0)	**0.9250**	0.2641	0.4006
b4	Overall	**0.1000**	0.3008	0.2041	c8	Overall	**0.9750**	0.1565	0.2970
	0(0.0)	**0.8950**	0.3073	**-0.5196**		0(0.0)	**0.0200**	0.1404	**-0.2644**
	1(1.0)	**0.1000**	0.3008	0.2041		1(1.0)	**0.9750**	0.1565	0.2970
b5	Overall	**0.1200**	0.3258	0.4414	c9	Overall	**0.9800**	0.1404	0.3412
	0(0.0)	**0.8750**	0.3315	**-0.6985**		0(0.0)	**0.0150**	0.1219	**-0.2307**
	1(1.0)	**0.1200**	0.3258	0.4414		1(1.0)	**0.9800**	0.1404	0.3412
b6	Overall	**0.0300**	0.1710	0.3987	c10	Overall	**0.9900**	0.0997	0.5100
	0(0.0)	**0.9650**	0.1842	**-0.5097**		0(0.0)	**0.0050**	0.0707	**-0.1353**
	1(1.0)	0.0300	0.1710	0.3987		1(1.0)	0.9900	0.0997	0.5100
b7	Overall	0.0300	0.1710	0.4207	c11	Overall	0.9750	0.1565	0.4225

	0(0.0)	**0.9650**	0.1842	**-0.5260**		0(0.0)	**0.0200**	0.1404	**-0.3750**
	1(1.0)	**0.0300**	0.1710	0.4207		1(1.0)	**0.9750**	0.1565	0.4225
b8	Overall	**0.0750**	0.2641	0.2153	c12	Overall	**0.9450**	0.2286	0.2582
	0(0.0)	**0.9200**	0.2720	**-0.4876**		0(0.0)	**0.0500**	0.2185	**-0.4652**
	1(1.0)	**0.0750**	0.2641	0.2153		1(1.0)	**0.9450**	0.2286	0.2582
b9	Overall	**0.0450**	0.2078	**0.1387**	c13	Overall	**0.9950**	0.0707	0.7417
	0(0.0)	**0.9500**	0.2185	**-0.3569**		0(0.0)	**0.0000**	0.0000	NaN
	1(1.0)	**0.0450**	0.2078	**0.1387**		1(1.0)	**0.9950**	0.0707	0.7417
b10	Overall	**0.1050**	0.3073	0.2104	c14	Overall	**0.9900**	0.0997	0.5100
	0(0.0)	**0.8900**	0.3137	**-0.5311**		0(0.0)	**0.0050**	0.0707	**-0.1353**
	1(1.0)	**0.1050**	0.3073	0.2104		1(1.0)	**0.9900**	0.0997	0.5100
b11	Overall	**0.0750**	0.2641	**-0.0661**					
	0(0.0)	**0.9200**	0.2720	-0.2553					
	1(1.0)	**0.0750**	0.2641	-0.0661					
b12	Overall	0.6300	0.4840	**-0.3524**					
	0(0.0)	0.3650	0.4826	**-0.2514**					
	1(1.0)	0.6300	0.4840	**-0.3524**					

Table 2 reveals that columns 1 and 2 provide classical item statistics (difficulty- p and discrimination- r_{pbs}) of local and foreign scales. On the CTT difficulty column (p) under the local intelligence scale ranges between 0.030-0.965 (3.0%-96.5%), while under the foreign intelligence scale, the difficulty index ranges between 0.000-0.99 (0.0%-99%). By implication, the range of examinees that got the items correctly under the local intelligence scale is between 3.0% and 96%. While under the foreign intelligence scale, examinees that got the items correctly ranges between 0.0% and 99%. Considering the criteria for the CTT difficulty index ($0.20 \leq p \leq 0.80$), items less than 0.20 are considered to be too difficult while those above 0.80 are too easy. Considering the difficulty index of foreign and local scale local appears more difficult than the foreign scale.

On the CTT discrimination index column(r_{pbs}) item discrimination values with poor discrimination power under the local intelligent test range between -0.066-0.138 (-6.6%-13.8%) while under the foreign scale ranges between -0.630- (-0.1353) (-63%- -23.5%) By implication the foreign intelligence scale have very weak discriminating power than the local test.

Research Question 3: Based on criteria set for the CTT framework ([a] $0.20 < P < 0.80$ and [b] $r_{pbs} > 0.15$) which and how many of the items survived under local and foreign intelligence tests?

To answer the research question, table 2 was reproduced in Table3. Looking at the left-hand columns of the table where CTT statistics are presented concerning the rule of thumb.

Table 3: Item Parameters Estimates and Survived Items

Item No	**Local Intelligent Test**					**Foreign Intelligent Test**			
	P	**Remark**	r_{pbs}	**Remark**		**P**	**Remark**	r_{pbs}	**Remark**
b1	0.4150	Good	0.2322	Good	C5	**0.9400**	Poor	0.4540	Good
b2	0.5850	Good	0.1577	Good	C6	**0.9950**	Poor	0.7417	Good
b3	0.1400	Poor	0.4107	Good	C7	**0.9250**	Poor	0.4006	Good
b4	0.1000	Poor	0.2041	Good	C8	**0.9750**	Poor	0.2970	Good

b5	**0.1200**	Poor	0.4414	Good	C9	**0.9800**	Poor	0.3412	Good
b6	**0.0300**	Poor	0.3987	Good	C10	**0.9900**	Poor	0.5100	Good
b7	**0.0300**	Poor	0.4207	Good	C11	**0.9750**	Poor	0.4225	Good
b8	**0.0750**	Poor	0.2153	Good	C12	**0.9450**	Poor	0.2582	Good
b9	**0.0450**	Poor	**0.1387**	Poor	C13	**0.9950**	Poor	0.7417	Good
b10	**0.1050**	Poor	0.2104	Good	C14	**0.9900**	Poor	0.5100	Good
b11	**0.0750**	Poor	**-0.0661**	Poor					
b12	0.6300	Good	**-0.3524**	Poor					

Table 3 reveals the assessment of the items using the set criteria for item difficulty, ($0.20 \leq p \leq 0.80$). Using these criteria, items whose difficulty index falls outside the range of 0.20 to 0.80 were considered poor. Columns 4 and 8 present the assessment of the items using the set criteria for item discrimination, $r_{pbs} \geq 0.15$ for local and foreign intelligent tests. Using these criteria, items whose discrimination index fell below or equal to 0.15 were considered poor.

Based on the stated criteria ($0.20 \leq p \leq 0.80$ and $r_{pbs} \geq 0.15$) for the classical item difficulty index, under the local intelligent test 9 items were considered poor while under the foreign intelligent test, 10 (all the items of the foreign scale) items were considered poor. By implication, the local intelligent test was too difficult for Nigerian testees ($p < 0.20$) while the foreign test was too easy for them ($p > 0.80$). Considering the classical discriminating index criteria under the local intelligent test 3 items were considered poor, while under the foreign test, all were considered good.

Research Question 4: How comparable are the surviving items from the local and foreign intelligence tests?

Table 4: Items deleted using the criteria set for CTT item parameters.

Item parameter	Number deleted	Item Deleted
Local Intelligence Test		
Difficulty	9	b3, b4, b5, b6, b7, b8, b9, b10, b11,
Discrimination	3	B12, b11, b9
	Western Intelligence Test	
Difficulty	10	C5,C6,C7,C8,C9,C10,C11,C12,C13,C14
Discrimination	0	0

Table 4 reveals that there is a variance in the number of items surviving from each of the intelligence tests. From the local test 9 items were found to have extreme difficulty indexes and 3 items had high discriminating indexes, while in the foreign test, with weak difficulty index 10 items showed extremely ranged values with no discriminating index score.

DISCUSSION

In the first research question, it was discovered that the local intelligent test reliability coefficient improved despite its weakness after the removal of item b12. The degree of skewness recorded by the two tests differs, the local test is mildly positively skewed while the foreign test is negatively skewed. By implication, the foreign scale appears too easy although with a kurtosis of about 72 score which shows many candidates got the test right. The mean score recorded after test scaling revealed that from the foreign intelligent test (mean= 0.971), 97.1% of the testees got the test right while on the local test (mean= 0.235), 23.5% of the testees got the test right. By implication, the foreign intelligent test is easier for Nigerian testees than the locally-made intelligence test. This result corroborated Culligan (2015) aimed to compare three common vocabulary test formats, the Yes/ No test, the vocabulary knowledge scale, and the vocabulary level test, as measures of vocabulary difficulty. The three tests were given to 165 Japanese students, the results indicated that the three tests measured one major latent trait (unidimensional) and they were significantly correlated in estimating their item difficulty.

On research question two, the result shows that the CTT difficulty column (p) under the local intelligence scale ranges between 0.030-0.965 (3.0%-96.5%), while under the foreign intelligence scale, the difficulty index ranges between 0.000-0.99 (0.0%-99%). By implication, the range of examinees that got the items correctly under the local intelligence scale is between 3.0% and 96%. While under the foreign intelligence scale examinees that got the items correctly ranges between 0.0% and 99%. Considering the criteria for the CTT difficulty index ($0.20 \leq p \leq 0.80$), items less than 0.20 are considered to be too difficult while those above 0.80 are too easy. Considering the difficulty index of foreign and local scales local appears more difficult than the foreign scale.

On the other hand, CTT discrimination index column (r_{pbs}) item discrimination values with poor discrimination power under the local intelligent test ranges between -0.066-0.138 (-6.6%-13.8%) while under the foreign scale ranges between -0.630- (-0.1353) (-63%- -23.5%) By implication the foreign intelligence scale have very weak discriminating power than the local test. This result corroborated Simbak, Aung, Ismail, Joush, Ali, Yaseein, Haque, and Rebuan (2014) who reported that the item discrimination index ranges between -1 and +1, the positive values are desirable, items with negative and zero values should be reviewed, zero discrimination indicates that the item does not differentiate between students. A negative and low discrimination index may result in miskeyed items or ambiguous items.

On research question three, the result was based on stated criteria ($0.20 \leq p \leq 0.80$ and $r_{pbs} \geq 0.15$) for the classical item difficulty index, under the local intelligent test 9 items were considered poor while under the foreign intelligent test 10 (that all the items of the foreign scale) items were considered poor. By implication, the local intelligent test was too difficult for Nigerian testees ($p< 0.20$) while the foreign test was too easy for them ($p> 0.80$). Considering the classical discriminating index criteria under the local intelligent test 3 items were considered poor, while under the foreign test, all were considered good.

This result corroborated Najar, (2010) and Alam, (2007) Item difficulty refers to the percentage of examinees who answered the item correctly, the values for item difficulty range from (0%-100%), and items with difficulty below 30% were considered to be difficult, while items with difficulty higher than 70% were considered to be easy. If the item has a low difficulty (less than 30%) there are several possible causes: the item may have been miskeyed, the item challenging the level of the student’s ability, or the item may be ambiguous. But, if the item has a higher difficulty (more than 70%), this could be explained by: the item being too easy, the item may have been miskeyed, or ineffective alternatives (Najar, 2010; Alam, 2007). Similarly, Simbak, Aung, Ismail, Joush, Ali, Yaseein, Haque, and Rebuan (2014) reported that to compare students’ performance on two evaluation techniques: multiple true-false and single best answer test formats, and correlated them with other assessment outcomes. The study analyzed the data for 20 item formats for each type of question, the participants were 3rd-year medicine students at Sultan Zainal Abdin University in Malaysia.

IMPLICTIONS

The result of the study suggests important theoretical and academic implications for research developers, psychometricians/psychological assessment specialists, scale development and implementation, that there are other factors outside the construct and establishing the psychometric properties of the scales that should be considered in the development of a scale, validating and revalidation of scales and implementation for psychometric adaptation, among which are the sample size and cultural settings of development and administration.

The study underscores the need for research institutes, and governmental and non-governmental agencies alongside scholars to get involved in the psychological assessment in designing and developing more statistical packages that are culture-based and user-friendly to enhance the diagnostic furtherance of educational and clinical intervention across Africa.

The study also underscores the need for engaging indigenous scales in more sophisticated revalidation procedures to engage more in the African intelligence test development. This study has also revealed that the indigenous standardized scale used in this study was difficultin terms of administration and this can reduce its psychometric strength thereby limiting its relevance to the ever-dynamic cultural proliferation as a result of societal evolution and advancement in technology. That is why the Western scale consistently showed superiority to indigenous scales because the foreign scale appears more organized and easy to administer.

LIMITATION

The study was limited to only Ibadan and did not consider other cities across Oyo states or Nigeria as a whole. Qualitative aspects such as interviews, observation, and focused group discussion (FGD) were not considered paramount as data was collected through a questionnaire only. The study was further limited by the unequal representation of both sexes and females participated more in the study than their male counterparts. The limitation of the study was also due to the unfamiliar procedure of administration and the items of the instruments were too many to answer following the adduced time recommended by the developer of Bakare Progressive Matrix, coupled with the language tone of the items which made the administration to be too intensive and rigorous. This study was carried out within a short period due to time constraints.

RECOMMENDATIONS

To boost the acceptance and adoption of indigenous standardized scales in the field of assessment and evaluation, the following recommendations were made based on the outcome of the study:

There is a need for psychological test developers to develop a well-standardized intelligence test that would be fair and culturally friendly among Africans and across the globe.

Efforts should be made by the government, research institutes, and educational bodies to ensure that the indigenous standardized scales are promoted and subjected to rigorous and sophisticated psychological testing that will promote replicative practices outside the shores of Africa that would be readily relevant, accepted, and practiced in Western communities.

Psychometricians, Clinicians, and Assessment experts should develop indigenous psychometric packages for validating psychological constructs that amplify the African culture to develop culture-based theories that relatively magnify its context as to native intelligence and personality. This will enhance the psychological coalition between Western and indigenous standardized intelligence scales. Educational bodies, Clinical and research institutes should ensure that they partner with governmental and non-governmental agencies in collaborating with scholars on developing and validating indigenous standardized intelligence scales within and outside the shores of Africa in other to improve the standards of indigenous scales to match the Western standardized scale such as Wechsler Adult Intelligence Scales(*WAIS-IV*), Divergent thinking Scale (*DTS*), etc., which have demonstrated a consistent psychometric strength and relevance even outside the Western world.

Counseling psychologists and Psychometricians should work on replicating Western psychological research about the African context most especially with the construct on personality and intelligence which should be culturally friendly and translated into the native language such as done by scholars in Australia, India, and Japan.

More funds should be put into research institutes to encourage scholars in engaging into a scholastic task that will be of benefit to the African context by developing programs that flag test bias on the intelligence test.

Government and non-governmental organizations should help create more favorable researchable platforms where indigenous scales are reviewed alongside the Western standardized scales by providing research grants.

Finally, there is a significant need to revise Bakare Progressive Matrix to maintain its psychometric integrity among other intelligence scales across the globe.

CONCLUSION

Although the study attempted to examine a comparison of indigenous standardized intelligence scales such as the Bakare Progressive Matrix and the Western standardized intelligence scale of the Wechsler Adult Intelligence scale to compare the intelligence of Adults. It is also important to note that a test is the only means to structurally authenticate the statutory construct, compare different scales, and content domain relevance of any research findings. It was discovered from the study that indigenous standardized intelligence scale such as Bakare Progressive Matrix was too difficult for the participants while the foreign test was adequate. This implies that there is a slight variance in the domain construct with timeline, considering the classical discriminating index criteria.

Though, this study focused on investigating test bias: a comparison between indigenous and Western standardized intelligence tests using the validation process of Bakare Progressive Matrix and Wechsler adult intelligence test as a case study carried out in Ibadan, Nigeria. It would be suggested that future researchers should try to focus on the following areas:

- There should be further collaborative studies on both indigenous standardized scales and Western standardized scales in other to promote the psychometric processes to match up standards with the Western scales.
- Extending the validity and reliability of indigenous standardized scales such as the Bakare Progressive Matrix, to a more sophisticated psychometric process to generate revised versions
- Research on Western standardized scales should be carried out within African settings to limit the tendency of cultural bias.
- There should be an extension of research in other areas such as indigenous intelligence scales that will in-cooperate theoretically cultural-based practices in Africa other than using the hybrid samples as perquisites participants such as African American, about psychological constructs (Intelligence).
- African scholars should develop a replicate prototype of all Western standardized Scales, such as done in India, Asia, Australia, Japan, Ghana, etc., to promote user-friendly scales relative to the African context.

REFERENCES

1. Aaron, Panofsky; Dasgupta, Kushan (2020). “How White nationalists mobilize genetics: From Genetic Ancestry and Human Biodiversity to counter science and Meta-politics”. American Journal of Biological Anthropology. **175**(2): 387–398.
2. Alam, S. (2007). Error analysis of Raven test performance. Personality and Individual Differences 16, 433–445
3. Bakare, C.G.M.(1977b). Study Habits Inventory (SHI), Student Problem Inventory (SPI), and Progressive Matrix. Ibadan: Psychoeducational Research Productions.
4. Braaten, Ellen B.; Norman, Dennis (2020). “Intelligence (IQ) Testing”. Pediatrics in Review. **27**(11): 403–408.
5. Brown, T 2019. Structural validity of the Bruininks-Oseretsky test of motor proficiency—Second edition brief form (BOT-2-BF). Res. Dev. Disabil., 85, 92–103
6. Culligan, B. (2015). A comparison of three test formats to assess word difficulty. Language testing. 32(4). 503-520
7. Culture Fair Intelligence Test using the Rasch model. Applied Psychological Measurement, 5355–5368.
8. Khan, S. (2019). A comparative analysis of emotional intelligence and intelligence quotient among Saudi business students toward academic performance. International Journal of Engineering Business Management, 11, 1847979019880665.
9. Mitchell, Kevin (2018). “Why genetic IQ differences between ‘races’ are unlikely: The idea that intelligence can differ between populations has made headlines again, but the rules of evolution make it implausible”.
10. Najar A.A (2010) A cross-cultural analysis of the fairness of the Cattell
11. Nitko, (2013). Cross-cultural normative assessment: translation and adaptation issues influencing the normative interpretation of assessment instruments. Psychological Assessment 6, 304–312.
12. Reynolds, C. R. (2000). Why is psychometric research on bias in mental testing so often ignored? Psychology, Public Policy, and Law, 6, 144–150.
13. Simbak, N. Aung, M. Ismail, S. Joush, N. Ali, T. Yassein, W. Haque, M. & Rebuan, H. (2014) Comparative Study of different formats of MCQs: Multiple true-false & single best answer test formats, in a new medical school of Malaysia. International Medical Journal. 21(6). 562-566
14. Stough, C.; Kerkin, B.; Bates, T. C.; Mangan, G. (2020). “Music and spatial IQ”. Personality and Individual Differences. **17**(5): 695.
15. S. 92011) Anxiety and test performance, ta. In: Spielberger, C.D., Vagg, P.R. (Eds.), Test Anxiety. Theory, Assessment, and Treatment. Taylor & Francis, Washington, DC, pp. 107–113.
16. UN Human Settlements Research Program, (2022). Report of the United Nations Human Settlements Programme (UN-Habitat) on human settlements statistics: note / by the Secretary-General
17. Wechsler, D. (2008). The measurement and appraisal of adult intelligence. Williams &Wilkins, Baltimore, 4 edition.
18. Whitaker, Simon (2020). “Error in the estimation of intellectual ability in the low range using the WISC-IV and WAIS-III”. Personality and Individual Differences. 48 (5): 517–521.
19. Van de Vijver, F. J. R. (1998). Towards a theory of bias and equivalence. *Zuma Nachrichten*, *3*, 41-65.