International Journal of Education and Social Science Research

ISSN 2581-5148

Vol. 6, Issue.3, May-June 2023, page no. 248-255

To cite this article: Kaku Zakka and Romy O. Okoye (2023). COMPARISON OF RELIABILITY COEFFICIENTS OF TESTS CONSTRUCTED USING DIFFERENT ORDERINGS OF DIFFICULTY OF TEST ITEMS, International Journal of Education and Social Science Research (IJESSR) 6 (3): 248-255 Article No. 788, Sub Id 1260

COMPARISON OF RELIABILITY COEFFICIENTS OF TESTS CONSTRUCTED USING DIFFERENT ORDERINGS OF DIFFICULTY OF TEST ITEMS

Kaku Zakka[1] and Romy O. Okoye[2]

[1]Department of education
University of Maiduguri

[2]Department of educational foundations
Nnamdi Azikiwe University Awka

DOI: https://doi.org/10.37500/IJESSR.2023.6321

ABSTRACT

The study compared reliability coefficients of tests constructed using different orderings of difficulties of test items for English language and Mathematics in the North-East, Nigeria. Experimental research design was adopted. The population of the study comprised all the 5,403 JSS 3 students of 2021/2022 academic session in the fifteen Unity schools of the zone. The sample used for experimental study consisted of three hundred and seventy-five students obtained through a combination of multi-stage and cluster sampling techniques. Two instruments (one for each of the subjects) were used for the study. Each instrument was prepared in three different formats, titled: Mathematic Achievement Test and English Language Achievement Test. Each format of the instruments was subjected to test-retest reliability. Method of data analysis was Pearson Product Moment Correlation coefficient (PPMC). The reliability estimate that were obtained from the mathematics multiple-choice questions are as follows: FORMAT ETD (0.512), DTE (0.849) and MDI (0.718). The reliability estimate that were obtained from the English language multiple-choice questions are as follows: FORMAT ETD ETD (0.728), DTE (0.669) and MDI (0.877). The study recommends that for English Language, items ordered from easy to difficult format should be used when constructing test items; since it was the one that had the highest reliability coefficient; for Mathematics items format from difficulty to easy should be used since it has the highest reliability coefficient among the test items format.

KEYWORDS: Reliability, Test item ordering, Order of difficulty, Test items, Test item arrangement.

INTRODUCTION

Tests are very essential in every educational institution. They are used in admission, placement, diagnosis, certification or ascertaining the extent to which educational objectives have been attained (Abanobi et al., 2023). Orluwene (2012) defined a test as an instrument used to determine the relative presence or absence of a trait being measured. Kapplan and Saccuzo (2001) defined a test as a measurement device used to quantify behaviour or an instrument which aids in understanding or predicting behaviour. This implies that there is an expected change in behaviour after instruction has

been administered and it is only through the use of a measuring device called test that such changes in behaviour could be ascertained and quantified. Iweka (2019) defined test as an instrument that is used to measure as accurately as possible the trait, character, personality or behaviour for which it is designed.

A test is not only used to assess human beings but also used to determine the relative presence or absence of a phenomenon, characteristic, attribute, feature or trait in a programme, school or textbook. A test as a measuring device can be used to holistically assess a learner. In this regard, Ukwuije (2007) defined a test as an instrument designed to measure the cognitive, affective and psychomotor components of an individual or a group of individuals. This means that tests could be used to assess the three domains of a learner. Hence, it is holistic in nature.

Test therefore can be defined as an instrument or a procedure for ascertaining the presence or absence of a trait or a behaviour in an individual or a group of individuals or a thing. When applied on human beings, students or learners, the aim is to determine a change in behaviour or prediction of behaviour in the cognitive, affective and psychomotor domains depending on the trait being measured.

Tests could be classified into essay (subjective) and objective tests. An objective test is a type of test administered to a leaner with the aim of assessing a particular aspect of the learner's knowledge by using questions which have one correct answer. This is opposed to a subjective test which has the aim of assessing areas of students' performance that are complex and qualitative by using questions which may have more than one correct answer or more ways to express it (Manuel, 2021). It therefore follows that in objective test, both the tester and the testee cannot influence the mark given. It is devoid of the feelings and personal interpretations of the scorers. In other words, if different examiners mark the same script of responses obtained from an objective test, using the same marking guide, they will come up with the same score (Okoye, 2015). A test must possess some fundamental qualities for it to be used as an instrument for measurement. These qualities mainly refer to its psychometric qualities which are reliability and validity, but the focus of the present study is on reliability.

Reliability simply refers to consistency in measurement. The reliability of a test is the consistency with which it yields the same result in measuring whatever it was designed for (Choudhurs, 2022). Reliability is the extent to which test scores are accurate and devoid of errors in measurement. A test is said to be reliable if it yields the same or similar scores over time, with repeated administrations of the same test or a parallel test.

Choudhurs (2022) discussed four main types of reliability namely: parallel forms, internal consistency, inter-raters and test re-test reliabilities. Choudhurs further explained that in parallel forms of reliability, the two different tests use the same content but separate procedures or equipment, and yield the same result for each test-taker. Internal consistency reliability is the type of reliability estimate in which items within the test are examined to see if they appear to measure what the test measures. The inter-

rater reliability makes use of two raters to score the same test and their inter-scorer consistency is determined. Test-retest reliability is a method of reliability estimate used when the same test is administered twice over a period of time usually between 7 to 14 days (Nwankwo, 2011; Nworgu, 2015; and Okoye, 2015).

A test item is a unit of questions or instruction that requires a testee or respondent to elicit a response. Several test items constitute a test. The University of Minnesota (2019) defined a test item as a specific task a test taker is asked to perform. A test item is used to assess a particular objective. More than one test item could also be used to assess a particular objective and such similar items could be grouped together to form subtests within a given test. These subtests or subscales in standardized tests are called test batteries. In other words, a group of test items addressing the same issue is simply referred to as a subtest or a test battery.

Investigate salient issues in students' poor performance in Mathematics in public Examinations in Nigeria: A case study of selected secondary school's student in Adamawa, North-East, Nigeria. Adopt expository research design, population of all students in SS 2 in Adamawa state, used stratified random sampling with 200 sampled. Finding revealed that in 2000 to 2009 academic session. In the year 2000, 4,942 sat for the NECO Examination in Mathematics and English in Adamawa state and only 190 students passed with credits that indicated 3.9% of the students, in 2001, 16,879 students sat for NECO in Adamawa North-East only 540 students passed with credit, in 2002, 7,188 students sat for Mathematics and English in NECO only 575 students passed with credit that indicate 3.0%. in 2009, 28,697 students sat for Mathematics in NECO and only 417 passed with credits while in 2014, 36,700 students sat for Mathematics and English in NECO only 2,600 candidates passed with credit which indicate 7.1% passed with credits. This showed the level of poor academic achievement of secondary schools' students in North East, Nigeria, Udonsu (2015).

The problem of poor result in national examination in Nigeria especially in the North East of the country has become a critical issue. High level of poor performance has been recorded especially in core subjects such as mathematics and English Language in the part of the country. In spite all efforts put up by government, schools, and other stakeholders, the poor results from student especially in the Basic Education Certificate Examinations, have persisted. It therefore becomes necessary to look at certain features of test used in conducting these examinations, with a view to ascertaining if they influence scores obtained with tests.

One of such features could be the way items of a test are ordered. There are different ways of ordering test items. They could, for example, be ordered following the sequence in which the topics are arranged in the scheme of work. Another way of ordering the items is to arrange them according to their difficulties. In this regard, they could be arranged in ascending or descending order of the item difficulty indices. The difficulty index of an item is a measure of the proportion of examiners who answered it correctly (Professional Testing, 2015).

International Journal of Education and Social Science Research

ISSN 2581-5148

Vol. 6, Issue.3, May-June 2023, p no. 248-255

Several works have been conducted on test item ordering and academic achievements of students. These include Opara and Uwah (2017) who investigated the effect of test item arrangement on performance in Mathematics among junior secondary school students in Obio/Akpor Local Government Area of Rivers State Nigeria. They discovered that item arrangement based on ascending order of difficulty had a positive and significant effect on students' performance in Mathematics while item arrangement based on descending order had a positive but non-significant effect on student' performance in mathematics. Similarly, Ollennu and Etsey (2015) worked on the impact of item position in Mathematics multiple-choice test on student performance at the Basic Education Certificate Examination (BECE) Level in Ghana and discovered a statistically significant difference in academic performance.

Chen (2012) carried out a study on the moderating effects of item order arranged by difficulty on the relationship between test anxiety and test performance and obtained the following findings: (1) the higher the test taker's level of test anxiety, the higher significance of the moderating effects and vice versa; and (2) item order adjusted according to individual examinee's perceived item difficulty may have a more significant moderating effect than item order arranged according to item bank calibrated item difficulty has. Marso (2022) conducted a study to determine if a relationship exists between test item arrangements and student performance on power tests and found significant influence of item ordering on academic performance. Schee (2013) conducted a study on the effects of item arrangements of multiple-choice examinations on students' performance and found out no significant effect of item ordering on the students' academic performance.

Most of the studies cited above, were concerned with effect of item ordering on academic achievement. It may be necessary to also ascertain how item ordering influences the reliability of the test. It is against this background that the researchers set out to compare the reliability coefficients of Mathematics and English language test constructed using different orderings of difficulties of test items.

Method: The experimental research design was adopted for the study. It was experimental because different groups of students were exposed to different experimental conditions.

Area of the Study: The study was carried out in the North-Eastern zone of Nigerian. The zone is made up of Adamawa, Bauchi, Borno, Gombe, Taraba and Yobe States.

Population of the Study: The population of the study comprised all JSS 3 students' of 2021/2022 session in all the 15 Unity Schools of the zone. According to records obtained from the Federal Education Quality Assurance Service (FEQAS), there were 5,403 JSS 3 students in the zone, during the 2021/2022 session.

Sample and Sampling Techniques: The sample used for the experiment consisted of 375 JSS 3 students obtained through a combination of multi-stage and cluster sampling techniques. At the first stage, three out 15 unity schools were drawn through simple random sampling. At the second stage,

three streams of JSS 3 were obtained from each of the three schools, through simple random sampling. All the students in the nine streams were used for the study. This gave rise to 375 students.

Instruments Used for the Study: Two instruments, each prepared in three different formats, were used for the study. They were titled: Mathematics Achievement Test and English Language Achievement Test. A draft of each of the instruments, which contained 80 multiple-choice test items, was first developed by the researchers. Each of these drafts was given to two experts in the corresponding discipline. The experts were requested to go through the items and assess them in terms of clarity of words, appropriateness of items and content coverage. Modifications were made in the items, based on the comments of the experts.

Each of the two drafts of the instruments was then administered on a sample of 150 students who were obtained from a school outside those used for the experiment. Thereafter, the difficulty indices of the items, for each of the tests, were computed using responses from the 150 students. Computations of the indices were done using a statistical software package named Item Analysis Statistical Package (IASP). Fifty items were then selected from the 80 items of each instrument. Only items having their difficulty indices ranging from 0.3 to 0.70 were selected. The fifty items constituted the test for each of the subjects.

The selected items for each instrument were then arranged in three ways, to give rise to three formats of the test. FORMAT ETD had the items arranged in increasing order of difficulty indices (Easy to Difficult- ETD). FORMAT DTE had items arranged in decreasing order of difficulty indices (Difficult to Easy- DTE), while FORMAT MDT had items arranged in random order of the indices (Mixed Difficulty- MDI). There were thus three formats of the test for each subject- DTE, ETD and MDI.

Method of Data Collection: With the test formats ready, the researchers proceeded to ascertain the reliability coefficient of each, for each of the subjects. The type of reliability sought was test- retest reliability. To achieve the above, in each of the schools sampled for the study, each of the test formats in Mathematics was randomly assigned to one of the three streams of JSS 3. Similarly, each of the formats in English Language was assigned to one stream, provided that no class was given the same format as in mathematics. The three formats for mathematics were administered in each school the same day by the class teacher, in form of ordinary class test. The formats for English Language were administered the next day by the English Language teachers. The reason for administering one single type to an in-tact class was to ensure that scores from the students would easily be matched with their scores during the next administration. For this reason, the students were requested to write their names on the scripts.

Fourteen days after the first administration, the same formats as in the first testing were administered on the respective streams, with Mathematics administered the first day and English Language the next day. The scripts of students, during the first and second administrations, were scored by the

researchers. Scores of students during the two administrations were matched, and the data subjected to statistical analysis.

Method of Data Analysis: The pairs of scores for each format were subjected to product-moment correlation analysis using the Statistical Package for Social Sciences (SPSS). This gave rise to the test-retest reliability coefficient for each format of the tests. These coefficients were then compared for each of the two subjects.

RESULTS AND DISCUSSIONS

Table 1. Reliabilities for Different Format for English Language

	ETD	DTE	MDI
Reliability	**0.728**	**0.669**	**0.877**

Table 1. shows the reliability indices of English Language of 0.728, 0.669 and 0.877 of test items ordering formats of Easy to Difficult, Difficult to Easy and Mixed Difficulty Index respectively.

TABLE 2. Reliabilities for Different Formats for Mathematics

	ETD	DTE	MDI
Reliability	**0.512**	**0.849**	**0.718**

Table 2. shows the reliability indices for Mathematics. The easy to difficult format had a reliability coefficient of 0.512. that of difficult to easy format was 0.849 while that of mixed difficulty indices format had reliability coefficient of 0.718.

DISCUSSION

Table 1:

From the result, it would be seen that for English Language, the MDI format had the highest reliability. This implies that arranging items in random order of difficulty makes the test to be reliable, when compared with the other formats. Specifically, one would say that scores of the students are most stable when the items are randomly ordered as against a situation when they are arranged from easy to difficult or from difficult to easy. This result was contrary to expectation. One had expected the reliability would be highest when arranged in ascending order considering the fact that tasks are easier to accomplish when one moves from simple to complex.

Table 2:

The result shows that the reliability of the test was highest when the items were arranged in decreasing order of difficult. This result differs from that of English Language. It is surprising that the result turned out as shown. Again, this finding went contrary to expectation, that the reliability should have been highest when the items are arranged from easy to difficult. This result might have occurred because of the nature of maths problems.

CONCLUSION

Based on the findings of the study, it is concluded that English Language, the reliability of a test is best when items are arranged in random order of their difficulty indices, while for Mathematics, the reliability is best when the items are arranged in descending order of their difficulty indices.

REFERENCES

Abanobi, C. C., Agu, N. N., Eleje, L. I., Metu, I. C., Mbelede, N. G., & Ezeugo, N. C. (2023). Effects of computer-based test (cbt) and paper and pencil test (ppt) on academic achievement and test anxiety of secondary school students' in economics. *Innovare Journal of Education*, 11(1): 35-41. DOI: https://dx.doi.org/10.22159/ijoe.2023v11i1.45960

Choudhurs, A. (2022). *Top 4 characteristics of a good test*. Retrieved March 24, 2022 from https://www.yourarticlelibrary.com/education/test/top-4-charact eristics-of-a-good-test/64804

Iweka, F. (2019). *Basic principles of educational measurement and evaluation.* Omoku: Chifas Nigeria.

Kaplan, R.M. & Saccuzo, D.P. (2001). *Psychological testing principles, applications and issues*. 5th Ed. Australia: Thomson Wadsworth.

Kpolovie, P.J. (2002). *Test, measurement and evaluation in education*. Port Harcourt: Emhai Printing and Publishing Co.

Manuel, J. (2021). *What are objective and subjective tests?* https://englishpost.org/objective-and-subjective-tests/

Marso, R. (2022). *Test item arrangement, testing time, and performance*. Retrieved March 24, 2022 from. https://onlinelibrary.wiley.com/doi/abs/10. 1111/j.17453984.1970.tb00704.x

Musari, A. (2016). *Much ado about educationally less developed states in Nigeria.* Retrieved March 24, 2022 from https://guardian.ng/features/much-ado-about-educationally-less-developed-states/

Nwankwo, O.C. (2011). *A practical guide to research writing for students of research enterprise*. (rev. ed.). Port Harcourt: Pam Unique.

Nworgu, B. G. (2015). *Educational research: Basic issues and methodology*. 3rd Ed. Nsukka: University Trust Publishers.

ISSN 2581-5148

Vol. 6, Issue.3, May-June 2023, p no. 248-255

Ollennu1, S.N.N. & Etsey, Y. K. A. (2015). The impact of item position in multiple-choice test on student performance at the basic education certificate examination (BECE) level. *Universal Journal of Educational Research 3*(10). Retrieved March 24, 2022 from https://files.eric.ed.gov/fulltext/EJ1077605.pdf

Okoye R.O. (2015). *Educational and psychological measurement and evaluation:* Awka: Erudition Publishers.

Opara, I.M. & Uwah, I. V. (2017). Effect of test item arrangement on performance in mathematics among junior secondary school students inObio/Akpor Local Government Area of Rivers State Nigeria, *5*(8). Retrieved March 24, 2022 fromwww.eajournals.org.

Orluwene, G. W. (2012). *Introduction to test theory and development process Port Hacourt:* Chris-Ron integrated Services.

Professional Testing (2015). Item analysis index. Retrieved March 24, 2022 from http://mededuunit.blogspot.com.ng/2015/07/the-difficulty-index-and-discrimination. html.

Reversa Dictionary (2022). *Test.* Retrieved March 24, 2022 from https://dictionary.reverso.net/english-cobuild/test+arrangement

Schee, B. A. V. (2013). Test item order, level of difficulty and student performance on principles of marketing education. *Journal of Education for Business*,14(23). Retrieved March 24, 2022 from https://www.researchgate.net/publication/271822453

Schee, B. A. V. (2022). Test item order, academic achievement and student performance on principles of marketing examinations. *Journal for Advancement of Marketing Education*,14(23). Retrieved March 24, 2022 from https://www.researchgate.net/publication/339103468

Ukwuije, R. P. I. (2007). *Appraisal techniques in guidance & counselling.* Port Harcourt: Chadik Press.

University of Minnesota (2019). *Test item.* Retrieved March 26, 2022 from http://psychology.iresearchnet.com/counseling-psychology/personality-assessment/psychometric-properties/