International Journal of Advanced Research in Public Policy, Social Development and Enterprise Studies
p-ISSN: 2536-6548 | e-ISSN: 2536-6556

IJARPPSDES

June, 2021
Vol. 4, No. 2

Psychometric Properties in English Language Exam: Case Study of NECO Exams 2014 – 2017

Theresa Baikwe Osusu
Faculty of Education,
Federal University Otuoke

Article DOI:
10.48028/iiprds/ijarppsdes.v4.i2.16

Abstract

An examination which is a Test constructed to cover areas to meet up summative Test NECO Examination is one of such bodies that have been noted for a cumulative sumature test construct it there for mil consist of all Psychometric properties need for a whole some Test that is qualified for its which are types validity and Reliability, item discrimination, item difficulty, guessing. Theories like Classical Test Theory (CTT), Latent Trait Theory (LTT) & Model of one Parameter Mode, two parameter model and three parameter logistics model of item response theory were all used empirical studies of this area were also stated. Two recommendations were also stated.

Keywords:
Psychometric properties, Validity, Reliability, Item discrimination and item difficulty

Corresponding Author:
Theresa Baikwe Osusu

Background to the Study

The determination of psychometric properties of tests is of great concern to researchers in both educational and psychological fields of study for several decades. Testing have been fully accepted in most societies as the most objective method of gathering information for decision making in school, industries, and government establishments. A test according to Ukwuije and Opera (2012) is defined as the administration of an instrument to the taste for determining achievement of some previously identified objectives in the individual. Orluwene (2012) regarded a test as an instrument used to determine the relative presence or absence of the trait measured. A test item is a set of interaction's collected together with any supporting materials and an optional set of rules for converting the candidate's response(s) into assessment outcomes. An item could be susceptible to guessing by examines during testing when it is not properly developed. Psychological testing deal with the development, validation, administration, coring and interpretation of psychological tests. Testing is used to evaluate, human abilities such as intelligence, aptitudes, skills, and achievement in various areas as well as personality characteristics which include traits, attitude, interest, skills, values or performance of individual or group. (Emaikwu, 2014). The basic challenge of educational system is concerned with the design and development of tests, the procedure for testing, instrument for measuring data and the methodology to understand and evaluate the results is therefore necessary that the quality of a test be of paramount importance to the test developer because it informs the decision to be made using the information provided by the test-item analysis of the examinees' responses to the items in a test is used to ascertain its quality by considering the psychometric properties.

Internal validity refers to whether the effects observed in a study are due to the manipulation of the independent variable and not some other factor. In other words, there is causal relationship between the independent and dependent variables. Internal validity can be improved by controlling extraneous variables. Using standardized instructions, counterbalancing, and eliminating demand characteristics and investigator effects. External validity refers to the extent to which the results of a study can be generalized to other settings (ecological validity), other people (population validity), and over time (historical validity), other people (population validity), and over time (historical validity). External validity can be improved by setting experiments in a more natural setting and using random sampling to select participants.

Assessing the Validity of the Test: There are two main categories of validity used to assess the validity of the test (i.e, questionnaire, interview, IQ test, etc.): Content and criterion.

a) **Face Validity**: Face validity is simply whether the test appears (at face value) to measure what it claims to. This is the least sophisticated measure of validity. Tests wherein the purpose is clear, even to naïve respondents, are said to have high face validity. Accordingly, tests wherein the purpose is unclear have low face validity (Nevo, 1985). A direct measurement of face validity is obtained by asking people to rate the validity of a test as it appears to them. This rater could use a Likert scale to

assess face validity. For example;

1a. The test is extremely suitable for a given purpose.
1b. The test is very suitable for that purpose
1c. The test is adequate.
1d. The test is in adequate.
1e. The test is irrelevant and, therefore, unsuitable

It is important to select suitable people to rate (e.g., questionnaire, interview, IQ test, etc). For example, individuals who actually take the test would be well placed to judge its face validity. Also, people who work with the test could offer their opinion (e.g, employers, university administrators, employers). Finally, the researcher could use members of the general public with an interest in the test (e.g., parents of testees, politicians, teachers, etc). The face validity of a test can be considered a robust construct only if a reasonable level of agreement exists among raters. It should be noted that the term face validity should be avoided when the rating is done by an "expert" as content validity is more appropriate.

b) **Construct Validity**: Construct validity was invented by Cronbach and Meehl (1955). This type of validity refers to the extent to which a test captures a specific theoretical construct or trait, and it overlaps with some of the other aspects of validity. Construct validity does not concern the simple, factual question of whether a test measures an attribute. Instead, it is about the complex question of whether test score interpretations are consistent with a nomological network involving theoretical and observational terms (Cronbach & Meehl, 1955). To test for construct validity, it must be demonstrated that the phenomenon being measured actually exists. So, the construct validity of a test for intelligence, for example, depends on a model or theory of intelligence. Construct validity entails demonstrating the power of such a construct to explain a network of research findings and to predict further relationships. The more evidence a researcher can demonstrate for a test's construct validity, the better. However, there is no single method of determining the construct validity of a test. Instead, different methods and approaches are combined to present the overall construct validity of a test. For example, factor analysis and correlational methods can be used.

c) **Concurrent Validity**: This is the degree to which a test corresponds to an external criterion that is known concurrently (i.e, occurring at the same time). If the new test is validated by comparison with a currently existing criterion, we have concurrent validity. Very often, a new IQ or personality test might be compared with an older but similar test known to have good validity already.

d) **Predictive Validit**y: This is the degree to which a test accurately predicts a criterion that will occur in the future. For example, a prediction may be made on the basis of a new intelligence test that high scorer at age 12 will be more likely obtain university degrees several years later. If the prediction is born out, then the test has predictive validity.

Reliability: Reliability refers to the consistency of a measure (Institute of Medicine, 2015). A test is considered reliable if we get the same result repeatedly. For example, if a test is designed to measure a trait (such as introversion), then each time the test is administered to a subject, the results should approximately the same. Unfortunately, it is impossible t calculate reliability exactly, but can be estimated n number of different ways.

a) **Test-Retest Reliabilit**y: Test-retest reliability is a measure of the consistency of a psychological test or assessment. This kind of reliability s used to determine the constituency of a test across time. Test-retest reliability is best used for things that are stable over time, such as intelligence. Test-retest reliability is measured by administering a test twice at two different points in time. This type of reliability assumes that there will be no change in the quality or construct being measured. (Leppink, Perez-fuster, 2017) state that in most cases, reliability will be higher when little time has passed between tests. The test-retest method is just one of the ways that can be used to determine the reliability of a measurement. Other techniques that can be used include inter-reliability, internal consistency, and parallel-forms reliability.

b) **Inter-Rater Reliability**: This type of reliability is assessed by having two or more independent judges score the test (Albers, 2017). The scores are then compared to determine the consistency of the raters estimates. One way to test inter-rater reliability is to have each rater assign each test item a score. For example, each rater might score items on a scale from 1 to 10. Next, you would calculate the correlation between the two ratings to determine the level of inter-rater reliability is to have raters determine which category each observation falls into and then calculate the percentage of agreement between the raters. So, it the raters agree 8 out of 10 times, the test has an 80% inter-rater reliability rate.

c) **Parallel-Forms Reliability**: Pararellel-forms reliability is gauged by comparing two different tests that were created using the same content. For Hu, Nesselroade, Ebrbacher, et al, (2017) says that is accomplished by creating a large pool of test items that measure the same quality and then randomly dividing the items into two separate tests, the two tests should then be administered to the same subjects at the same time.

d) **Internal Consistency Reliability**: this form of reliability is used to judge the consistency of results across items on the same test (Institute of Medicine, 2015). Essentially, you are comparing test items that measure the same construct to determine the test internal consistency. Because the two questions are similar and designed to measure the same thing, the test take should answer both questions the same, which would indicate that the test has internal consistency.

There are a number of different factors that can have an influence on the reliability of a measure. First and perhaps most obviously, it is important that the thing that is being measured be fairly stable and consistent (Polit, 2014). If the measured variable is something that changes regularly, the results of the test will not be consistent. Aspects of the testing situation can also have an effect on reliability. For example, if the test is administered in a room that is extremely hot, respondents might be distracted an unable to

complete the test to the best of their ability. This can have an influence on the reliability of the measure. Other things like fatigue, stress, sickness, motivation, poor instructions, and environmental distractions can also hurt reliability.

Item difficulty: Item difficulty expresses the proportion or percentage of students. Who answered the item correctly? Item difficulty parameter measures the difficulty of answering the item correctly and it ranges from 0.00 (none of the student answered the item correctly) to 1.00 (all of the students answered the item correctly). Several factors could contribution to the difficulty index of an item in a given test such as content/learning objectives was not taught, distracters are not clear enough; maybe more than one likely answer, student in general did not come prepared for the exam, syllabus too expensive and students were unable to cover or revise the whole prescribed course, incompetent teacher such he is unable to deliver the learning objectives wholly and properly (Nauman, 2016). These factors could dispose the students to guessing in the examination either increasing or decreasing their score. Guessing should be discouraged by means of instructions give on the test and by scoring the test in such a way as to penalize those who guess incorrectly by the use of formula scoring.

Item Discrimination: Item discrimination index indicates how well the in term serves to discriminate between student with higher and lower levels of knowledge. The discrimination parameter is a measure of the differential capability of an item usually denote by a high discrimination parameter value suggests an item that has a high ability to differentiate subjects. In practices, a high discrimination parameter value means that the probability of a correct response increases more rapidly as the ability (latent trait) increases.

Guessing: This means giving an answer or making a judgment about something without being sure of all the facts. Guessing is a serious problem in examinations (Obinne, 2012) which usually affects traits to ability estimation of the examinees. Traits or skills are grouped into cognitive, effective and psychomotor skills. These test properties are used to estimate a person's ability or latent trait, these traits or constructs are measured by drawing items that sample observable aspects of the constructs or traits based on a particular theory or model. Guessing is seen by many students and teachers as a major factor that determines the scores of an examinee in an objectives test, guessing is a serious problem that must be dealt with. Guessing increases measured error score and it raises possibility of correct responses. The uses of guessing tactics and strategies introduces error that weakens the relationship among items in term of their psychometric properties, which will adversely affect the trait or construct under investigation. However, to an examiner, guesswork appears to be the only available way to increase his score in a given test, guessing is a skill, based on critical analysis of options presented in an item seen as intelligent guessing. In the bid to control the effect of guessing, correction factor for guessing was introduced as one of the items leading to guessing parameter in item response theory, which is currently incorporation un the three-parameter model of the IRT. However, the challenges to test developer are to draw item of a test not prone to

guessing by students. Therefore, test developers must ensure that test items to examine should be constructed to minimize both error in construction and item guessing by the examinees so as to ensure that the test is both valid and reliable especially in the English language subject. Most test administered to student based on classical test theory where guessing was not adequately controlled constitute serious problem in the determination of the ability of students. Guessing seems to be most preferred option when mathematics items are based on critical thinking or reasoning. Obinne, (2011) found out in that both National Examination Council (NECO and West African Examination Council (WAEC) biology test items of the year 2000 contain biased items whose c-valued (guessing parameter) lie above 0.30 which is very high.

Classical Test Theory (CTT): This has been the foundation for measurement theory for decade and this theory revolve around three main concepts. Test score (often called the observe score), True score and Error score. The classical test theory suggest that any assessment will only reveal an individual observed score and that this is not always a reflective f their "true score" as there is something (error) in the environment that impacts on the individual performance. Different errors are associated with measurement theories which includes random errors, systematic errors, errors associated with test and errors associated with the testees guessing test items, each of these errors must a great extent be controlled I a good testing process.

The model is given as:
$X = T + E$

Observed scores (X) – this is the score obtained by individual on an assessment. True score (T) – this is the individual true ability and is always constant for a particular person. Error € - anything that may have impacted an individual's performance on a test, either increasing or decreasing factors such as guessing by examine during testing.

The major advantages of CTT are that is can be performed with smaller representatives' samples of examinees and again its application is very easy in term of analysis of its data. Classical test analysis employs relative sample methodical procedures hence commonly used by most classroom teachers in assessing students. The potential problem with CTT method is sample untested assumption that the items within a test are inter-chainable in contributing to a total test score, that is the latent ability estimate is the total score and measurement error is assumed constant across he trait level which is problematic for making comparisons across different test forms, other limitation of CTT include the fact that item parameters in CTT is sample-dependent, it has parallel test form issues and therefore can be problematic in making comparison across different test forms or compare examinees' scores, CTT lacks predictability since error is assumed the same for everybody. The major aim of CTT within psychological testing is to understand and improve the reliability of psychological tests and assessment. Reliability of a test in the CTT is determined by the correlation coefficient between the observed scores on to parallel measurements, as the reliability of a measurement increase, the error variance becomes

relatively smaller (Adedoyin, 2010). When the error variance is relatively small, an examinee's observed score produced a rater poor estimate of the true scores. However, item statistics (item difficulty and item discrimination) are also an important part of ICTT mode (fan, 1998). Theoretically item response theory overcomes the major weakness of CTT, hence it is widely preferred option.

Latent Trait Theory

Latent Trait Theory otherwise known as the item Response (IRT) evolved doe to the weakness of classical test theory by providing a reporting scale on which examinee ability (the construct measured by the test) is independent of the particular choice f test items that are administered, it implies that examinee's endorsement of an item does not affect his next choice of another item on the same test the term latent is used to emphasize that discrete item response are taken to be observable manifestation of hypothesize traits, construct por attributes not directly observed, which must be inferred from the manifest response. Measurement error are controlled in latent trait theory in line the model involved especially in three parameter model where a parameter is assigned to guessing which is the focus of the work.

Latent trait model is seen as in improvement over classical test theory, it is more sophisticated and allows for the improvement of the reliability of an assessment. It's emphasis is confused on three notions; a unidimensional trait denoted by 0, local independence of items and item response function (IRF) or item characteristics of both test takers which it is exposed and other items that constitute the test in general, it is statistical theory about examinee's item and test performance and hoe performance relates to the abilities that are measured by the items in the test, it provides more adaptable and effective method of test constructions, analysis and scoring than those derived from classical test theory or model. A model can represent a theory in the sense that it interprets the laws and axioms of that theory. In IRT there exists different parameter models adjusting for different item properties leading to different ability estimation. These models include:

1. One-parameter model (also known as the Rasch model) which adjust for item difficulty level as the trait level required to correctly answer a question.
2. Two-parameter model (2PL) which accounts for item difficulty and discrimination parameters.
3. Three-parameter Logistics model Item Response theory (IRT) developed by Lord, (1980). The model has basic assumptions of un-dimensionality and Local Independence. In the 3-parameter Logistic model, the probability of a correct response to a dichotomous item. Usually, a multiple-choice item is presented mathematically as follows;

$$Pi(\theta) = Ci + \frac{1 - Ci}{1 + e - Dai(\theta - bi)}$$

Where:

Pi: is the test taker's ability
Ai: is the item discrimination index.
Bi: is the difficulty parameter
Ci: is the guessing index
E: is the base of natural logarithm and approximately equal to 2.714
D: is the arbitrary constant (normally D = 1.7)

IRT which is also known as latent response theory is the probability of answering an item correctly or of attaining a particular response level in relation to individual ability and characteristics of the item. The goal of IRT is to predict the probability at which a testee f a given ability level responds to an item correctly. In IRT ability level is measured on a transformable scale having a mid-point of zero, a unit measure of one with the theoretical range of ability from negative infinity to positive infinity, however, practical consideration usually limits the range of values from -3 to +3 (Hambleton et al., 1991). According to Zaman et al. (2008). The ability range in IRT estimates is between – to + theoretical but typically they range from + 3.0 for examples with high abilities on the test to -3.0 for examples with low abilities. The difficulty estimates in IRT for items range from +3 to -3 the item with difficulty level + 3 and -3 are labelled as "very difficult) and "very easy" respectively. There are them IRT models for test items that are dichotomously scored known as three, two and one-parameter IRT model to describe the items the parameter is a-parameter (discriminating power), b-parameter (difficulty level) and c-parameter (guessing factor). The value of item difficulty denoted by b-parameter is a location parameter that indicate the position of the item characteristic curve in relation to the ability that is required for a testee to have 50% chance of getting the item right the item discrimination denoted by a-parameter provides information on how will an item separate testee with high and low ability level while guessing factor denoted by c-parameter indicated the ability at which testee guess answer correctly that is the effect of guessing on the probability of a correct response. The value of these parameter indicates the ability level at which they occur which practically ranges from -3 t0 +3. IRT provides a framework for evaluating how ell individual item in a test or examination function. IRT enables the psychometricians to develop and design examination items, maintain item banks, and equate the difficulties of item for successive version of examination which allow comparison between result overtime. According to Yu (2008), IRT address the weakness of Classical test Theory (CTT). CTT does not provide information about how examinees at different ability level perform on the item. IRT is a necessary tool which has to be at any testing centres for a valid instrument (Tshering, 2006).

According to Adedoyin (2010), for more objective educational measurement, IRT theoretical framework should be incorporated by examination bodies in Africa for the construction of examination items. Three-parameters model (3PL). this considers the effect of item guessing in addition to the difficulty and discrimination levels of the item. This model assumes that the three parameters, difficulty, discrimination, and guessing are combined for an estimate of a relationship between the probability of a correct of an item

and the trait level (ability) of an examiner. The three different parameter models may yield different ability estimates. However, other factors that may affect the estimate of the ability include the dimensionality of the test and test scoring format (dichotomous or polytomous). Three parameter model (3PL) considers the combination of the difficulty, discrimination and guessing parameters.

This study was guided by the following empirical works which were conducted by some researchers in the related area. A study by Fehintola, & Akingbade (2019) investigated the comparison of psychometric properties of multiple-choice test using confidence and number right scoring among Senior Secondary School students in Ibadan metropolis. The study adopted a descriptive design of survey type. The population for the study consisted of Senior Secondary School Two (SSII) students in Ibadan Metropolis, Oyo State, Nigeria. A sample of 400 Agricultural science students was selected across 4 level Local Government Areas in Ibadan metropolis, using purposive (mainly Agricultural Science Students) and random sampling techniques. The instrument used for the study was Agricultural Science Multiple-choice Test. The 50 items Agricultural Science 4-option test was administered on the student. Data collected were analyzed using paired samples t-test, Kuder-Richardson (KR-21), Cronbach alpha, and Fisher z-test. The results obtained revealed that significant difference existed in the difficulty indices with Number Right (NR) and Confidence Scoring Method (CSM) with mean of 55.42 and 44.01 respectively. Also, there was a significant difference in the CSM and NR in the discrimination indices with NR and CSM has mean of 0.57 and 0.52 respectively. It was found that NR significantly improved the difficulty and discrimination indices. Furthermore, the finding revealed that there was no significant difference between NR and CSM in the reliability coefficient. Based on these findings, it was recommended that number right scoring method would be used to assess Agricultural Science Student's performances because it makes test item appears moderate in terms of difficulty level and is very easy for student to guess the item right.

According Psychometric analysis of Senior Secondary School Certificate Examination (SSCE) 2017 NECO English Language Multiple Choice Test Item in Kwara State Using Item Response Theory was conducted by (Jimoh, Aina, & Akintomide, 2022). They determined the dimensionality of 2017 National Examinations Council (NECO) English Language multiple-choice test item and estimated the item parameter indices (discriminations, difficulty, guessing and carelessness) using four parameter logistic model. The ex-post facto design was employed for the study. The population for the study comprised all candidates/test-takers who enrolled and sat for June/July Senoir School Certificate Examination (SSCE) 2017 NECO English Language Examination in Kwara State, Nigeria with 12,000 samples purposively selected from sixteen Local Government Area in the State. the research instruments used for the study were optical Marks Records Sheets for the NECO June/July 2017 English Language objectives items. The responses of the testees were scored dichotomously. The data collected were calibrated using four parameters logistic model. The results showed that the 2017 English Language multiple-choice item among SSCE student in Kwara state does not violate the assumption of

unidimensional which made the items reliable for use in assessing knowledge of student in English Language. Also, the result showed that only two items were able to suit the 4-PLM based on the rule of thumb. While the remaining items does not suit the 4-PLM. It was recommended among others that NECO and other examination bodies should intensify more efforts improving the standard of the English Languages test items using 4-PLM, which is the new trend for estimating item parameter indices.

Obinne (2011) conducted a study on the psychometric analysis of the two major examinations conducted in Nigeria by NECO and WAEC. The objective was to compare the standard error of measurement of Biology examinations conducted from 2000 – 2002 using the one-parameter model of Item Response Theory (IRT). Standard error of measurement (EM) is commonly used to produce confidence interval and it is an estimate of how much error there is in a test. Instrumentation research design was used for this study. Benue State, Nigeria was the study area. the population for the study comprised all year three (SSIII) Senior Secondary School students who enrolled for May/June/July 2006 Biology Senior Secondary School Certificate Examination of NECO and WAEC in the three education zones of Benue State. The sample for the study was one thousand eight hundred (1800) student. Multi-stage stratified sampling techniques of the BILOG MG Computer Programme and the SPSS were used for data analysis. The results whose significant differences in the SEM of Biology examinations conducted by NECO and smaller SEM (high reliability) than those of WAEC). It has recommended that IRT analysis should be employed by Nigerian Examination bodies.

The study by Ogunbamowo (2019) determined the dimensionality of WAEC and NECO Economics test items and assessed the difference in each of item discrimination, difficulty, and the guessing parameter of the two tests as obtained using CTT and IRT. These were with the view of determining the comparability of the two examinations under different test theories. The research design adopted for the study was descriptive. The population for the study consisted of secondary school student in Osun State and a sample of 540 students. The instruments used for the study were adopted respectively from the 2017 Economics WAEC and NECO Senior School Certificate Examination Titled Economics Achievement test 1 (EAT 1) and Test 2 (EAT 2). The results showed that the difference in the discrimination indices of NECO and WAEC Economics test items when CTT as used is not significant ($U = 1.52$, $P > 0.05$). However, there is a significant difference ($U = 3.029$, $P < 0.05$) in the discrimination indices when IRT was used. The results also showed that while the difference in difficulty indices of NECO and WAEC Economics test items was not significant with the use of CTT ($U = 0.138$, $P > 0.05$), the difference was significant when IRT was used ($U = 2.095$, $P > 0.05$). the results further showed that difference in the guessing parameter of NECO and WAEC Economics test items is not significant ($U = 1.519$, $P > 0.05$). The results concluded that while the two examinations were comparable under classical test theory, they are not comparable under item response theory. Public examinations such as the National Examinations Councils (NECO) in Nigeria are established to provide a leverage ground for all examines to obtain the Senior School Certificate. The test items from this examination body are expected not to favour any

group examinees, either from urban or rural, private of public, male or female, or other groups. It is also expected that the psychometric properties of the examination items should be such as should be able to distinguish between the best-able and least-able examinees.

Conclusion

This study is a resume introducing the work to a pour of direction to an analysis of NECO Examination of 2014 – 2017. Examination is to be guided by all psychometric test properties in this work. How and what was done in all examinations will be considered and this will lead to conclusion on the work was through.

Recommendations

1. All external examinations like NECO and others should make that all Psychometric Properties need to be checked and given to measurement experts before they are administered.
2. English Language is compulsory subject and compulsory as an ending requirement therefore all psychometric properties should be in place.

Reference

Adedoyin, R. (2010). *Investigating the variance of the person parameter estimate*, hltps://www.research.gatenet.

Albers, T. (2017). *Quantitative data analysis is- in the graduate curriculu*m, https:// journal sagepub.com

Cronbach, W. & Meehl, V. (1995*). Construct validity in psychological tests*, https:// psyche ap. Org.

Fan, J. (1998). *Direct estimation and linear component for high dimension data*, Princeton University, https://Fan princeton.edu. publication.

Fehintola, G. & Akingbade, U. (2019). *Comparison of psychometric properties Al-hkmah University Ilorin*, https ://www alhikmah.edu.ng

Hambelton, et al (1991). *Fundamentals of items response theory*, Whelan.

Jimoh, et al (2022). *Constructs to serve as basis semantic scholar*, http://www.semasntic.scholar.com

Lord, F. M. (1980). *Application of item respondent theory to practical testing problems 1st Edt* Routlege, http://wwwtarylorfrancis.com

Nauman, S. (2010). *Lack of critical thinking skills leading to research crisis in developing countries: A case of Pakistan*, Research Gal.

Tshering, U. (2006). *Educational research reviews comparative analysis of test construction academia*.edu.

Vevo, B. (1985). Concept of face validity, *Journal of Educational Measurement & Evaluate*. https://psyenet.apa.org