British International Journal of Education And Social Sciences
Vol.7, No.11; November- 2020
ISSN (3342 – 543X);
p –ISSN (4519 – 6511)
Impact factor: 8.11

APPLICATION OF ITEM RESPONSE THEORY ON DEVELOPMENT OF ACHIEMENT TEST IN SECONDARY SCHOOL ECNONMICS IN ENUGU EDUCATIONAL ZONE, ENUGU, NIGERIA

[1]Bernadine Amarachi Ugwuanyi and [2]Genevieve Chimaoge Ebulum
[1]Department of Science & Vocational Education (Measurement & Evaluation), Godfrey Okoye University (GOU), Enugu.
[2]Department of Psychology, University of Nigeria Nsukka
Corresponding author: Genevieve Chimaoge Ebulum**,**

Abstract: The study aimed at applying the Item Response Theory (IRT) on the development of achievement test in Secondary Schools in Enugu Education Zone. Four research questions guided the study. The study adopted instrumentation research design. The population for the study was 3105 SS2 students. A sample size of 1015 was obtained using multiple stage random sampling technique. The instrument for data collection of this study was Economics Achievement Test (EAT) which consists of 60 test items, face validated by three specialists, one from Economics Department and two from Measurement and Evaluation unit; all from Godfrey Okoye University (GOU), Enugu. The reliability coefficient of 0.77 was obtained using Kuder – Richardson 20 (K-K$_{20}$). A total of 1005 out of 1015 copies were correctly filled and retrieved immediately and were used for the study. The data collected on the study were analyzed using BILPG – MG for IRT analyses including probability level, item goodness of fit and others. The results from findings revealed that 95% of the multiple choice test in Economics had high reliability in measuring the students' ability. Fifty eight percent (58%) of the multiple choice test in Economics fitted the two parameter (2PL) model because the items were equal and above 0.05 level of significance. Educational Implications of the findings of the study show that: IRT is a reliable measure and essential to the field of Economics as well as, to the study of latent abilities of students, IRT is a valid methodological framework for modeling response data from assessments in Economics, A didactic application of IRT highlights some of these advantages for psychological test development. Often, classical methods exclusively are used to offer evidence of validity and reliability to new tests and this evidence is undoubtedly important. But the findings of this study provide evidence that IRT results can be extremely helpful to complement this evidence with information regarding the quality of the measurement at specific points of the measuring scale.

Keywords: Item Response, Achievement Test, Economics

INTRODUCTION

Achievement test is a test prepared by teacher in order to measure the extent the students have attained on a particular topic or course which has been exposed to them previously. Ani (2014) stated that achievement test is an examination designed to assess how much knowledge a person has in a certain area or set of areas. Also, from birth to old age, a person encounters tests at almost every turning point in life. Psychological tests sum up performance in numbers or classification. Tests measure individual differences in traits or characteristics that exist in some vague sense of the word.

Objective test is one of the assessments used in testing or assessing student's academic achievement in any given instruction. In objective tests, such as multiple choice questions, students are asked and respondent required to selects the best possible answer(s) out of the choices from a list (Phabean, 2017). The multiple choice items consist of a stem and a set of options which examiners can choose from, with the correct answers and incorrect answers.

Test scores obtained from multiple choice questions are used to assess the competence of students. Multiple choice test, unlike essay test allow the teachers to ask a large number of questions that adequately cover the course content (Okoro, 2009). All assessment instruments must satisfy the criteria of reliability and validity (Jerome, 2013).

Reliability is concerned in relation to extent of consistency or dependability of a measuring instrument (Abonyi, 2011). This implies that if any test were to be applied in Economics in an infinite number of times, it would be expected to generate responses that vary a little from trial to trial, as a result of measurement error. Hence, for any measuring instrument, the smaller the error, the greater the reliability while the greater the error, the smaller the reliability.

Validity refers to the extent to which an instrument measures what it is designed to measure (Nworgu, 2015). A test with high validity will measure accurately the particular qualities it is supposed to measure. Usability of test is the extent to which a test provides the teacher or test administrator clear instructions that can be put into practice without a great deal of difficulty or confusion. In other words, a test in economic is useable if it doesn't force students to waste their time dealing with the idea of recording the answer. Nonetheless, instrument development in Economics requires more than determination of reliability, validity, and usability of the items. Some other indices such as item difficulty, item discrimination, and item distractor are required for of the quality of the measurement. With regards to the requirements of the National Policy of Education, teachers are expected to construct and administer valid and reliable instrument with a view to determine the extent of attainment of educational objectives by their students.

In Nigerian secondary schools, it appears that Economics teachers are failing to consider the issue of validity and reliability in the assessment and measurement procedures required for the test items in Economics. Quite often, validity is sacrificed for reliability and this usually results in measurement being only concerned with the precision of scores rather than the intellectual values. Based on the above observations, there is need for the development and provision of good quality instrument to enable teachers take decision on integration of the individuals into a sound and effective citizen and provision of equal education opportunities for all citizens as gender equality and also to measure the different latent abilities among them (gender)

Gender refers to biological and social differences between men and women. According to Ukagwu (2018), gender refers to the socially, culturally constructed characteristics and roles which are ascribed to males and females in society. The influence of agenda on students achievement in Economics is still controversial and inconclusive. For instance, Adewolu (2013) stated that gender had no significant in Economics. This finding was contradicted by Ani (2014) who stated that girls achieved higher than boys in Economics achievement test. This study seeks to contribute in resolving this controversy by introducing the development of achievement test as a multi-step process that can allow one of the two distinct measurement frameworks which are called the classical test Theory (CTT) and item response theory (IRT).

According to Obinne (2008), IRT is a body of theory that describes the application of mathematical models to data from questionnaires and tests as a basis for measuring things such as abilities and attitudes. Item response theory is referred to as the strong true score theory or modern mental test theory because, it is a more recent body of theory that makes stronger assumptions as compared to classical test theory (Emaikwu, 2012). This approach to testing based on item analysis considers the chance of getting particular items right or wrong. Item response theory as a body of theory provides a basis for evaluating how well assessments work in terms of being able to assess the individual ability. In education, psychometricians apply IRT for examinations and equating of the difficulties of successive versions of examination. One of the advantages of IRT according to Ani (2014) is that it provides a measure of precision of ability estimate to each ability level. Thus, instead of providing a single standard error of measurement that applies to all examinees, that's, provides a separate estimate of error for each examinee and each of it. Due to the fact that classical test theory has failed to provide satisfactory solutions to many testing problem like test equating, test bias and test development, there is a great need for using item response theory which deals with item parameter and person parameter for the development of instrument which will be used to assess the item characteristics in relation to the individual's ability.

Research questions

1. What are the standard errors of measurement of the test items of the multiple choice test in economics?
2. How does the Economics achievement test item fit the two parameters logistic (2PL) of IRT models?

METHODS

This study is instrumentation in nature. Instrumentation according to Ayozie (2015) is research design which is planned for study which enables the researchers develop and often validate instrument required for execution of prescribed tasks in education. It is instrumentation in nature since it is used in construction, validation and population of valid and reliable test in other to assess the students' achievement in Economics. The sample for the study consists of one thousand and fifteen (1015) SS2 Economics students. The instrument consists of 60 multiple choice questions with four (4) options (A-D).

Content validation of the test was carried out by preparing the table of specification based on the six levels of cognitive domain of Bloom's taxonomy of education**,** the EAT was administered to twenty five (25) SS2 Economics students.. Their responses were scored and analyzed, while Reliability coefficient of 0.77 was obtained using Kuder-Richardson 20 (K-R$_{20}$) formula to determine the internal consistency of the instrument.

The data for this study were collected through the use of Economics Achievement test (EAT). The researcher visited the sample schools to collect the data for the study. The copies of the instrument were administered to the students through the assistance of Economics teacher in the respective sampled schools

The data for this study were collected through the use of Economics Achievement test (EAT). The researcher visited the sample schools to collect the data for the study. The copies of the instrument were administered to the students through the assistance of Economics teacher in the respective sampled schools.

The tests were administered to the students under a good atmosphere and the test lasted for 50 minutes. Maximum likelihood estimation technique of the BILOG – MG V3 of two parameter logistic (2PL) MODEL Computer programming was used to answer the research questions. BILOG – MG is compatible for Item Response Theory (IRT) analysis of dichotomously scored items data, including probability level, item goodness of fit and differential item functioning

Results

Research Question One: What are the standard errors of measurement of the test items of the multiple choice test in Economics?

Table 1: Standard errors of measurement of the test items of the multiple choice test in Economics based on two-parameter logistic (2PL) model.

Item	S.E	Item	S.E	Item	S.E	Item	S.E	Item	S.E	Item	S.E
1	0.09	11	0.15	21	0.08	31	0.14	41	0.04	51	0.44
2	0.24	12	0.09	22	0.03	32	0.05	42	0.55	52	0.09
3	0.07	13	0.05	23	0.08	33	0.08	43	0.07	53	0.13
4	0.15	14	0.11	24	0.07	34	0.33	44	0.16	54	0.44
5	0.10	15	0.09	25	0.08	35	0.06	45	0.05	55	0.27
6	0.10	16	0.36	26	0.15	36	0.08	46	0.09	56	0.12
7	0.16	17	0.08	27	0.07	37	0.06	47	0.10	57	0.09
8	0.06	18	0.06	28	0.05	38	0.22	48	0.07	58	0.61
9	0.61	19	0.10	29	0.07	39	0.14	49	0.16	59	0.09
10	0.11	20	0.08	30	0.08	40	0.12	50	0.20	60	0.08

Decision rule: The lower the SEM, the higher the reliability.

The result in Table 1 indicated the standard errors of measurement of the test items of the multiple choice questions in Economics based on two parameter logistic (2PL) model. This shows that majority of the items except items 9, 42 and 58 have a standard error of 0.03 to 0.44. This shows that fifty seven (57) items representing (95%) had standard error below 0.50 and three (3) items representing (5%) have standard error above 0.50. Hence, the smaller the error, the greater the reliability while the greater the error, the smaller the reliability. This finding is consistent with those of Zanon, Yoo, Hutz (2016). This high reliability indicated consistency in measuring the students' ability in Economics.

Research Question Two: How do the Economics achievement test items fit the two parameters (2PL) of IRT models?

Table 2: Fits statistics of Economics multiple choice test based on two parameter logistic (2PL) model.

Item	Chi.sq	Prob	Item	Chi.sq.	Prob	Item	Chi.sq.	Prob	Item	Chi.sq.	Prob
1	35.2	0.06	16	39.3	0.07	31	51.3	0.09	46	42.5	0.07
2	77.8	0.08	17	18.0	0.12	32	29.3	0.06	47	45.5	0.12
3	23.8	0.05	18	31.7	0.01	33	23.2	0.04	48	94.3	0.03
4	70.4	0.06	19	44.2	0.00	34	179.9	0.08	49	43.7	0.09
5	102.5	0.07	20	77.4	0.09	35	51.2	0. 02	50	21.3	0.05

6	26.1	0.03	21	13.7	0.02	36	23.2	0.00	51	93.4	0.00
7	41.0	0.06	22	92.6	0.04	37	149.9	0.14	52	37.6	0.08
8	18.4	0.00	23	45.4	0.24	38	116.5	0.05	53	67.3	0.02
9	76.0	0.13	24	18.0	0.14	39	26.1	0.03	54	48.5	0.12
10	43.9	0.10	25	79.0	0.02	40	41.0	0.09	55	51.9	0.00
11	79.4	0.05	26	46.0	0.02	41	13.7	0.04	56	47.6	0.14
12	57.1	0.01	27	35.2	0.01	42	77.8	0.12	57	96.6	0.02
13	31.5	0.13	28	31.4	0.02	43	52.1	0.06	58	30.5	0.12
14	18.2	0.11	29	84.2	0.13	44	31.7	0.01	59	90.9	0.01
15	55.0	0.04	30	48.7	0.08	45	44.2	0.00	60	46.7	0.03

Decision Rule: Reject HO: iff $\chi^2_{cal} > \chi^2_{cuv}$ at $\alpha = 0.05$, otherwise accept. Since $\chi^2_{cal} = 58\% > \chi^2_{cuv} = 42\%$, we do not accept.

Table 2 revealed the chi-square goodness-of-fit analysis for the items of the multiple choice questions in Economics based on two parameter logistic (2pl) model. The criterion for all the items fit/misfit was determined at 0.05 level of significance. Results revealed that the chi-square value linked with the probability value ranged from 0.00 to 0.29. Based on the data in table 2, thirty five (35) items representing (58%),

Discussion

The findings of the study are discussed in line with the research questions that guided the study. Specifically, the study was discussed according to the following sub-headings:

- Standard errors of Measurement of the test items of the multiple choice test in economics.
- Fitness of economics achievement test items to the two parameters (2PL) models.

Standard Errors of Measurement of the Economics Multiple Choice Test

This study discovered standard error of measurement of 60 test items of multiple choice questions in Economics. Findings revealed that 57 Items : 1-8, 10-41, 43-57, 59 and 60 (95%) have a standard error of 0.03 to 0.44 while items 9, 42 and 58 have standard error above 0.50.Standard error above 0.50 is described as low reliability. Hence, the smaller the error, the greater the reliability while the greater the error, the smaller the reliability. This finding is consistent with those of Zanon, Yoo, Hutz and Humbletom (2016) that IRT test information functions provide the amount of information or measurement precision captured by the test on the scale measuring the construct of interest and other features too.

These findings of IRT application on multiple choice Economics test provide different standard errors of measurement at different trait levels. Because standard errors of measurement are used in score interpretations, it is possible to easily create confidence intervals to interpret individual scores (Embretson, 1996). So, one can have a range (around the reached score) associated with a probability. The smaller the errors at some level, the smaller the confidence bands. This does not happen with classical methods that nearly always assume the same standard error applies at all trait levels.

Fitness of Economics Achievement Test Items to the Two Parameters (2PL) Models

Findings of this study revealed that the chi-square value linked with the probability value ranged from 0.00 to 0.29. Based on the data in table 2, thirty five (35) items representing (58%), and 58 fitted the two parameter model because some of the items were equal to and some were above 0.05 level of significant. While twenty five

(25) items representing (42%) did not fit the two parameter model because the items were below 0.05 level of significant. These characteristics are possible because IRT models provide item and ability parameter invariance for test items and persons, when the IRT model of interest actually fits the available test data. In other words, the same items used in different samples will keep their statistical properties (for instance, difficulty and discrimination), and persons' scores that represent ability or latent traits on a specific construct will not depend on the particular test items they were administered. To gauge how well the chosen model can predict respondent scores and generate item statistics that are invariant over samples of respondents, it is essential to measure model fit. This involves fitting the Graded Response Model (GRM) to the data, and for estimating item and latent trait parameters.

Conclusion:

The study aimed at applying the Item Response Theory (IRT) on the development of achievement test in Secondary Schools in Enugu Education Zone. Four research questions guided the study. The study adopted instrumentation research design. The population for the study was 3105 SS2 students. A sample size of 1015 was obtained using multiple stage random sampling technique. The instrument for data collection of this study was Economics Achievement Test (EAT) consist of 60 test items, face validated by three specialists, one from Economics Department and two from Measurement and Evaluation unit, all from Godfrey Okoye University (GOU), Enugu. The reliability coefficient of 0.77 was obtained using Kuder – Richardson 20 (K-K_{20}). A total of 1005 out of 1015 copies were correctly filled and retrieved immediately and were used for the study. The data collected on the study were analyzed using BILPG – MG for IRT analyses including probability level, item goodness of fit and others. The results from findings revealed that 95% of the multiple choice test in Economics had high reliability in measuring the students' ability. Fifty eight percent (58%) of the multiple choice test in Economics fitted the two parameter (2PL) model because the items were equal and above 0.05 level of significance. Educational Implications of the findings of the study shows that: IRT is a reliable measure and essential to the field of Economics as well as, to the study of latent abilities of students.

Recommendations:

1. IRT should be used to improve the measurement of scales with much precision
2. Test makers should use knowledge of the features of IRT to refine and increase the validity and reliability of other psychological measures
3. IRT produces person parameter invariance (test scores are not dependent on the particular choice of test items) when model fit is present. Hence, the topic of model fit should be used by test makers to prove important characteristics of IRT analysis in development of psychological tests.

REFERENCES

Adebule, S. O. (2012). Item Response theory as a basic for measuring latent trait of interest. Greener *Journal of Social Science, 3* (7), 378 - 382.

Abarghoie, H. M., Khamiripoor, Y. M., Hosseini, H., Esmaeili, B., & Abarghoie, J. M. (2012). Development and standardization of achievement test. *Journal of American Science,* *8*(4).

Adegoke, B. A. (2013). Comparison of item statistics of physics achievement test using classical test and item response theory frameworks. *Journal of Education and Practice 4*(22).

Ani, E. N. (2014). *Measurement and Evaluation for Teacher Education (2nd ed.*). Enugu: SNAP Press ltd.

Anime S. (2014). *Dynamics of Economics*. Available at https:www.fandim.com,accessed 1/4/9.

Ani, E. N. (2014). Application of item response theory in the development and validation of multiple choice test in economics.

Ayozie, D.O. (2015). Overview of measurement and evaluation. *Journal of Social Sciences,* 3 (a), 204 -211.

Asadu, I. N. (2010). Trend in student's enrollment and performance in senior secondary certificate examination in economics. *Unpublished doctoral dissertation.* University of Nigeria Nsukka.

Awopeju, O. A. , & Afolabi, E. R. I. (2018). Comparative analysis of classical test theory and item Response theory based item parameter estimates of senior school certificate mathematics examination. *European Scientific Journal, 12*(28)

Barker, F. B. (2011). *The basics of item response theory, (5th ed*.). United State of America. ERIC clearing house on assessment evaluation.

Bichi, A. A. (2013). Item Analysis using a derived science achievement test data. *International Journal of Science and Research, 4(*5).www.ijsr.nets

Birnabum, A. (2010). *Statistical theory for logistic mental test models with a prior distribution of ability.* Princeton, NJ: Educational Testing service.

Callaliam, P. (2012). *When and How to Apply Item Response Theory* . New Jessey: Eugine Press.

Egunjibi, A., & Egwaukhide, F. (2010). *Economics for Senior Secondary School.* Lagos: MaCMilliam Nigeria Publishers Ltd.

Emaikwu, S.O. (2011). Issues in test bias public examinations in Nigeria and implementations for testing. *International Journals of Academic Research in Progressive Education and Development, 1*(1) (pp.40).

Nenty, H. I. (2004). From classical test theory (CTT) to item test theory (IRT) : An introduction to a desirable transition. In: O. A Afemikhe, & J. G. Adewale (Eds) : Issues in educational measurement and evaluation in Nigeria. *Institute of Education, University of Ibadan, Nigeria,* PP. 372-384.

Nworgu, B.G. (2015). Introduction to Education Measurement and Evaluation: Theory and Practice (2nd ed). Nsukka: Hallman Publisher.

Obinne, A.D.E. (2008). Psychometric Properties of Senior Certificate Biology Examination Conducted by West African Examination Council: Application of item Response Theory. Unpublished Doctoral Dissertation, University of Nigeria Nsukka

.

Obiune, A.D.E. (2013). *Using IRT in Determining test item prone to guessing.* Reprieved June, 20, 2013, URL. https.//dx.doi.org/wje.ve.

Okeke, F.N. (2008). Women and Leadership in his/her education; facing international challenges and maximizing opportunities. Association *of Common Wealth University Bulletin, 147,* 14-17.

Okoro, C.O. (2010). Development and validation of extracurricular instructional package in social studies. Faculty of Education University of Port Harcourt, Port Harcourt River State Nigeria. *Journals of Academia* Retrieved May, 30, 2013, from http://www.science pub.net.

Obinne A.D.E. (2014). Test item validity: item response theory (IRT) perspective for Nigeria. *Research Journal in Organizational Psychology & Educational Studies 2* (1). Retrieved journal, 28, 2014, fromwww.emergingresource.org.

Osadebe, P. U. (2014).Construction of economics achievement test for assessment of students. *World Journal of Education, 4*(2).

Palmieri, P. A. (2012). Item Response Theory method and application gaining support as assessment

Sial , Z. A. and Khan, K. M. (2012). Development of an achievement test in the subject of health and physical education at intermediate level in Pakistan. *Journal of Elementary Education, 24*(1), 61-70.

instrument. Retrieved December, 18, 2012, from http://www.istss.org/publication.

Phebean, A. (2017). *An item response theory analysis of the academic: A motivation inventory for secondary students in South Western Nigeria.* Availahle at www.academic.edu. accessed on 2ndFebruary 2019.

Reve, B.B. (2000). Item and scale-level analysis of clinical and non-clinical sample responses to the MMPI-2 depression scales employing item response theory. Unpublished doctoral dissertation, university of North Carolina at chapel Hill.

Sharma, J. (2012). *Principles of Application of item Response Theory*. New York: Achullan.