Test Dimensionality and Conditional Independence of Undergraduate Physics Test Items of National Open University of Nigeria: Perspective from Latent Traits Model (Ltm) Package of R Language

*Nadrah Nadrah**

[1]*Faculty of Teacher Training and Education, University of Muhammadiyah Makkasar, Indonesia; E-mail: nadrah@unismuh.ac.id*

Abstracts: Physics test items developed by the National Open University of Nigeria (NOUN) to measure test-takers' proficiency in Elementary Mechanics, Heat, and Properties of Matter (PHY101) are intrinsically multidimensional. However, we observed the current method used by the University to score test-takers proficiency in the course is premised on Classical Test Theory, which has been criticised in the literature for its limitations, and a measurement model for unidimensional tests. Consequently, using an inappropriate model to model test-takers responses to items would adversely affect their true proficiency in the course. Therefore, this study assessed the dimensionality and conditional independence of physics test items of NOUN using the ltm package in the R language. A non-experimental design of survey research type was adopted. Test-takers responses to the 35 PHY 101 multiple-choice items across NOUN study centres in the 36 States of Nigeria were retrieved from the Directorate of Examination of Assessment (DEA) and used for the study, with empirical reliability of 0.80. Nine hundred seventy-eight test-takers responses were captured and analysed with modified parallel analysis test and Yen Q3 statistics implemented in ltm package of R language software, version 4.0.2. Findings remarked that NOUN physics items had more than one predominant dimension to account for the observed performance of test-takers in the course. Also, residual correlations of ten item pairs severely violated the conditional independence benchmark of 0.20. We conclude that PHY 101 of NOUN is intrinsically multidimensional. Test-takers responses to the test items are due to their locations on multiple latent variables. We recommend that a psychometric unit be established within the DEA of the University to check the appropriateness of the test items developed by the faculty lecturers.

Keywords: Physics Test-Items; Item Response Theory, National Open University of Nigeria, Conditional Independence, Dimensionality, Latent Trait Model, R Language, Modified Parallel Analysis Test

1. INTRODUCTION

One of the most crucial factors that affect human education and thinking is the curriculum they are taught with its important elements and distinctive skills and types. Therefore, curricula specialists seek to introduce contemporary issues and skills into textbooks to play an important role in achieving sustainable human development. All kinds of thinking skills are among the basic skills that contemporary education seeks to develop among learners. Modern educational trends emphasize the importance of providing learners with thinking skills and practicing them, and perhaps visual thinking skills represent one of the thinking forms that has received wide attention in the educational area [1].

Globally, the hallmark of every academic endeavour at all levels of education is attaining such an educational program's outlined aims and objectives. The most productive tool for determining the appropriate level of attainment is through assessment of the learners via examination [2]. Brink and Lautenbach (2011) further posited that one of the pivot pillars of measuring the success or otherwise of any academic institution is the degree of quality and successes recorded in the conduct of her examination and assessment. The National Open University of Nigeria (NOUN) was established in 1982 on open and distance education philosophy. Her mandate is to offer quality higher education and increase capacity building to the highly growing population of Nigerian seeking higher education while remain actively engaged as working-class citizens. As open and distant learning (ODL) institution that is riding on the advancement in information and communication technology [3], NOUN has extended across the 36 States of the country with an impressive capacity to deliver credible examination across her 102 study centres (Author) [4].

Examination and Assessment is an integral part of academics embedded in the teaching and learning process [5]. An examination is a systematic process of measuring learning outcomes after a specified period. It allows faculty members to measure student achievement of learning objectives towards taking valuable and informed decisions regarding the academic growth of learners, improvement in instructional modalities, and the institution in general [6]. The successful conduct of semester examinations often begins with developing quality test items. Quality items are essential for building credibility, validity, and reliability in the examination process towards drawing valid conclusions from the resulting scores. However, research studies have revealed that the development and validation of quality test items, with specific reference to multiple-choice questions having appropriate psychometric properties, are usually inundating [7].

In NOUN, the task of examination and assessment is shouldered by the Directorate of Examination and Assessment (DEA). It is saddled with the responsibilities of conducting all examinations and handling assessment-related matters, including the Tutor Marked Assignment (TMA). The semester summative examinations at NOUN have been broadly classified into multiple-choice questions (MCQ), which are rendered through electronic examination (e-exam), and the essay type commonly referred to as Pen-on Paper (PoP). By the University policy, an e-exam is designed for 100 and 200 level students in all the faculties except students in the faculty of Law. The traditional assessment method of Pen-on Paper (PoP) is reserved for all students in 300 levels and above, including the postgraduate students. In addition, the semester TMA form the basis of the students' continuous assessment test is primarily an e-exam. This huge responsibility of examination and assessment is always in collaboration with the academic departments and faculties that make up the University, and with strong technical support from the Management Information System (MIS), the Directorate of Information and Communication Technology (DICT), and the Learning Content Management System (LCMS) [8]. All NOUN's academic and non-academic organs highlighted above are working together to ensure the quality delivery of electronic examinations and assessments.

Presently, the Head of Departments (HODs) across the eight different Faculties in the University – Education, Science, Health Sciences, Management Sciences, Social Sciences, Art, Agriculture, and Law – are saddled with the responsibility of developing the examination questions by their respective lecturers for onward delivery to DEA in preparation for semester examination. Although, the submitted questions are often subjected to several quality assurance checks from the originating departments during internal and external moderation exercises before getting to DEA. The standard practice of setting e-exam questions for all 100Level and 200Level courses requires that the lecturers handling each of the courses ensure the development of 35 standard items for MCQ and FBQ, respectively, on all assigned courses. For a given course in the first and second year of any program, a total of 70 items must be ready for e-examination [9].

However, all the highlighted processes and procedures are insufficient to guarantee quality assured test delivery in the modern measurement community. There are gaps to be bridged in the overall procedures of developing the test items to meet international best test development practices. Measurement theories such as Classical Test Theory (CTT) and Item Response Theory (IRT) are designed to develop quality, valid, and reliable test instruments. The CTT measurement models and procedures for developing educational tests and interpreting test scores have been the foundation theory in the measurement parlance and served testing experts adequately for decades [10]. Many tests have been developed using this approach of Classical models and procedures, and NOUN is not exempted from using this model to develop their multiple-choice items for 100 and 200 levels students. For, author documented some of the shortcomings that characterised adoption of CTT, such as (a) use of item parameters whose values depend on the particular group of students with which they are obtained and (b) person parameter estimates that depend on the particular choice of items selected for a test.

Due to limitations evident in using CTT to develop quality test items, psychometricians and measurement specialists have advanced a new measurement system, called item response theory (IRT), to address these and other limitations of common educational measurement practices. IRT is a general framework for specifying mathematical functions that describe the interactions of persons (examinees) and test items based on underlying ability [11]. As of today, IRT is essentially used commonly by the largest testing companies in the developed clime such as the United States and Europe for the design of tests, test assembly, test scaling and calibration,

development of test item banks, assessments of test item bias and other common procedures in the test development process. It is noteworthy that measurement communities, researchers, and public school systems have endorsed and employed IRT with increasing enthusiasm and frequency in the developed countries to developing assured quality test items. This assertion is in the light of the paradigm shift from the CTT to IRT. More importantly, before a researcher could opt for this modern measurement theory, there are a set of assumptions required to establish the IRT model's data. However, the practicability of assumptions cannot be ascertained directly [12]. Some indirect evidence can be assessed, and the model's overall fit to the test data can be evaluated. These assumptions are dimensionality, conditional independence, and monotonicity or item response function (IRF).

2. DIMENSIONALITY ASSUMPTION OF IRT

A common assumption of IRT models is that only one ability or trait is measured by a set of items in a test. In reality, this assumption is difficult to strictly met. Various cognitive skills, personality, and test-taking factors (such as level of motivation, test anxiety, ability to work quickly, tendency to guess when not sure about answers, etc.) can mar the test performance at least to some extent (De Mars 2010). The expectation is that the test items should measure with a single dominant factor that influences students' test performance for uni-dimensionality. The dominant factor is the ability measured in the test items, which is not necessarily intrinsically. With the intervention of learning, forgetting, etc., ability scores vary over time. Summarily, when one factor accounts for the test performance of a student in a test is referred to as a uni-dimensional model. Also, Models in which it is assumed that more than one trait is required to account for student test performance are referred to as multidimensional [13].

Multidimensional item response theory (MIRT) is another case of IRT built on the premise that the mathematical function includes a vector of multiple person characteristics. This describes the skills and knowledge that the person brings to a test and a vector of item properties that explains the difficulty of the test item and the sensitivity of the test item to differences in the characteristics of the individuals [14]. However, these multidimensional models are more complex to date, and researchers in sub-Saharan Africa are just beginning to embrace their development and usage for test items (Author). For instance, a test comprises collections of test items. Each of the test items is difficult in its way. The students who respond to the test items are of various diverse individuals. Even children born on the same day will show some differences in their knowledge and skills because their life experiences are different after birth. Students' interactions with the test items on a test score in a set of responses represent complex processes. To display the complexity of a test item and clarify the components of a test item. Let's consider one of the NOUN physics 101 multiple-choice items. This test item is a measure of physics.

For a student to select a response option for this test item, it is the product of interaction between their capabilities and the characteristics of the test item. This test item requires different types of knowledge and several skills to get the correct answer “B”. Thus, the number of factors responsible for the observed variation in student scores in the test item is multidimensional [15].

Several methods that can be used for assessing dimensionality of a binary scored data, such as the non-linear method implemented in the program called DIMTEST package for Stout’s test of essential dimensionality (De Mars 2010; Finch and French 2015), full information of item factor analysis (FIFA) implemented in (TESTFACT, EQSIRT, and MIRT R Package) DETECT index of dimensionality, Mokken scale analysis, and Normal Ogive harmonic analysis robust method (NOHARM) (Fraser and McDonald 2003 cited in Adewale et al. 2017). In this study, the Bootstrap Modified Parallel Analysis Test (BMPAT) implemented in the Itm package in R language was used to assess the dimensionality of NOUN physics 101 test items [16]. Several researchers such as Yu et al. (2007); Finch and Monahan (2008); Sijtsma and van der Ark (2017); Okwilagwe and Ogunrinde (2017); Zhang and Stout (2016); Author; Author; Emons, Sijtsma, and Pedersen (2012); Bastug (2016); Author; have used various methods to establish dimensionality of dataset either dichotomous or polytomous in their various studies. Results from their studies remarked a unidimensional factor explaining the variance observed in students' test scores (Finch and Monahan 2008; Sijtsma and van der Ark 2017; Zhang & Stout 2016; Author; Author). However, studies conducted by Emons, Sijtsma, and Pedersen (2012); Li, Jiao and Lissitz (2012); Bastug (2016); Oguoma, Metibemu, and

Okoye (2016); Author asserted that many large scale achievement tests consisted of multiple content areas or domains within a single subject. These areas or domains caused multidimensionality or local item dependence, which both violate the assumptions of the unidimensional IRT models currently used in many large-scale assessments. NOUN, which is the focus of this study, is not exempted [17].

3. CONDITIONAL INDEPENDENCE ASSUMPTION OF IRT

Conditional or local item independence emphasises the assumption that students' independent responses to all test items when their abilities (θ) are conditioned. This is used to show that responses are assumed independent at the level of individual students with the same value of θ. Still, the assumption does not generalise to the case of variation in theta. Fienberg and Linden (2018) [18] assert that in the groups of individuals with variation in the trait being assessed, responses to different test items typically are correlated because they are all related to levels of the individuals' traits. This implies that the conditional independence assumption is the student's response to one item does not affect their response to another item. This is met in uni-dimensional IRT models when the probability of a student's response pattern is equal to the product of probabilities associated with the student's score for each item [19].

More so, any time a set of test items refers back to the same stimuli, this assumption is threatened. When this assumption is violated, trait estimates may be inflated due to overestimating item information (Author). Hambleton, Swaminathan, and Rogers (1991), cited in Author, present a mathematical expression of the definition of conditional independence as:

$$P\left(U_1, U_2, ..., {U_n}/{\theta}\right) = P\left({U_1}/{\theta}\right), P\left({U_2}/{\theta}\right), ..., P\left({U_n}/{\theta}\right)$$

$$For \quad P_i(\theta) = P\left(U_i = {1}/{\theta}\right) and\ Q_i(\theta) = P\left(U_i = {0}/{\theta}\right)$$

$$P\left(U_1, U_2, ..., {U_n}/{\theta}\right)$$

$$= P\left({U_1}/{\theta}\right), P\left({U_2}/{\theta}\right), ..., P\left({U_n}/{\theta}\right)$$

$$= P_1(\theta)^{u_1} Q_1(\theta)^{1-u_1}, P_2(\theta)^{u_2} Q_2(\theta)^{1-u_2}, ..., P_n(\theta)^{u_n} Q_n(\theta)^{1-u_n}$$

$$= \prod_{i=1}^{n} P_i(\theta)^{u_i} Q_i(\theta)^{1-u_i}$$

…… Equation 1

Where θ is the ability assumed to influence the student's ability on the test, U_i is the response of a randomly chosen student to item i (i=1,2,…,n), and $P\left(U_i = {1}/{\theta}\right)$ is the probability of correct while $P\left(U_i = {0}/{\theta}\right)$ is the probability of an incorrect response.

There are several methods for assessing the assumption of conditional independence as recommended by [20]. These are; likelihood ratio G2 (C and T G2; Chen and Thissen 1997), the power-divergence (PD) statistic, the Q3 statistic (Yen 1984), Fisher's r-to-z transformed Q3 (Yen 1984; 1993), the Wald test, the likelihood ratio test in logistic regression (LR G2), Chen Thissen LD, the absolute value of mutual information difference [21], the mutual information difference (MID), the modification index (MI) in structural equation modeling (SEM), and the use of the residual correlation from the factor analysis (FA). However, Yen (1984; 1993) suggested that the best method for assessing the conditional independence assumption is Yen's Q3 statistic, and this method was used in this study.

4. MONOTONICITY ASSUMPTION OF IRT

The monotonicity assumption takes the form of the normal ogive. The item response curve has a mean of 0 and a standard deviation greater than 1. This is also known as item response function (IRF) or item characteristic curve (ICC). An ICC is a mathematical function that relates the probability of success on an item to the ability measured by the test containing the items. Bichi and Talib (2018) noted that the ICC remains invariant from one group of students to the next, resulting in the invariance of item parameters involved in generating the ICC. This is an important aspect of the modern theory which distinguishes it from the CTT.

In this study, Multiple-Choice Questions in Elementary Mechanics, Heat, and Properties of Matter (PHY,101) were used to examine the framework that is in operation at NOUN. PHY101 is a 2 credit unit course and one of the numerous courses taken at 100 Level by all students in the Faculty of Science and others in the Department of Science Education. The existing negative perception about physics courses is yet to be erased in the mind of so many students who have nurtured unfounded fear about the subject (Author; Author). Adopting such a subject as a case study for this research is not out of place. It is a way of tackling what may appear as a twin challenge for the subject [22].

Physics test items developed by NOUN measured students' proficiency in the course, which are inherently multidimensional. However, we observed that the current method used by the University to assess student's ability in the course is premised on CTT implicit in the uni-dimensional model (Author). Consequently, using the inappropriate model to model students' responses to test items would adversely mar their true proficiency in the course. The practice of test development without determining the dimensionality and conditional independence of such test items has become an outdated practice in measurement parlance [23]. The modern approach situated in the IRT has remained highly unpopular among test developers, of which NOUN is one. To the best of our knowledge and evidence from literature, no study has been conducted to establish the IRT assumptions of dimensionality and conditional independence of NOUN 100level physics test items, thereby the study filling the gap.

Consequently, the study poised to answer the following questions: Do NOUN Physics test items obey the assumption of the IRT framework of dimensionality? Do NOUN Physics test items follow the assumption of the IRT framework of conditional independence and monotonicity, respectively? The study's findings will contribute to the existing discussion in the literature regarding the establishment of IRT assumptions for test items in large-scale assessment [24].

5. METHOD AND PROCEDURES

5.1. Research Design, Participant, and Procedure

A non-experimental design of survey research type was adopted. The target participants for the study were undergraduate students offering PHY101. The census sampling technique, which captured all the PHY101 students, was adopted for the study. This sampling technique was considered appropriate since all the population under concern was studied. Student responses to the 35 PHY 101 multiple-choice items across NOUN study centres in the 36 States of Nigeria were retrieved from the Directorate of Examination and Assessment (DEA). This body is responsible for all examinations and assessments matters in the University. The instrument consists of topics such as a sample of test item reads as "The path followed by the projectile is", "A jet lands on an aircraft carrier at a speed of 63 m/s. What is the magnitude of its acceleration (assumed constant) if it stops in 2.0 s?", "Magnitude of a vector quantity can best be represented by", and "Materials which lengthen considerably and undergo plastic deformation until they break are called", and so on. Each item is scored as 1 – correct or 0-incorrect [25]. The secondary data gathered were used for the study, with empirical reliability of 0.80. Nine hundred seventy-eight test-takers responses (370 females and 608 males) were captured and analysed using modified parallel analysis test, likelihood ratio test, and Yen Q3 statistics implemented in ltm package of R language software, version 4.0.

6. STUDY RESULTS AND DISCUSSION

The responses of test-takers on physics 101 across the 36 States, including the federal capital city in Nigeria, were used to conduct this analysis. The analysis assessed the dimensionality and conditional independence of PHY 101 test items administered by the National Open University of Nigeria (NOUN). To achieve this feat, test-takers responses were subjected to latent trait model (ltm) package implemented in open-source R programming language for statistical computing version 4.0.2. Tables 1-2 and Figure 1 present the modified parallel analysis test and plot.

Table 1: Modified Parallel Analysis Test for NOUN physics Test Items

	Monte Carlo samples	T	p-value
Second eigenvalue in the observed data	100	1.41	
Average of second eigenvalues in Monte Carlo samples		0.99	0.06

Table 1 depicts the dimensionality result of 978 responses of test-takers to physics 101 test items. The study hypothesised that the second eigenvalue of the observed data is not substantially more significant than the second eigenvalue of simulated data under the assumed item response theory (IRT) model was therefore rejected with (T = 1.41, 0.99, p = 0.06). The difference observed between the second eigenvalue in the observed data, and the random generation eigenvalue is significant. This implies that NOUN physics 101 test items are inherently multidimensional. The observed responses on these test items had more than one latent variable that reflected test-takers proficiency (Ɵ). To further ascertain the result from modified parallel analysis, one-dimension and two-dimensions were hypothesised using the likelihood ratio test implemented in the ltm package. Table 2 present the result.

Table 2: Likelihood Ratio Test for Dimensionality of NOUN physics Test Items

Dimension	AIC	BIC	log.Lik	LRT	df	p-value
One- Dimension	38207.3	38378.29	-19068.65			
Two- Dimension	36986.6	37499.58	-18388.3	1360.7	70	<0.01

Table 2 presented results when one and two-dimension were hypothesised. The result showed statistics for Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), the Log-likelihood ratio for both hypothesised dimensions. The result yielded a significant value at two-dimensions with (LRT = 1360.7, df = 70, p = 0.01). Also, it was remarked that AIC, BIC, and log-likelihood values for two-dimensions were lesser than one-dimension, which corroborate findings (See Table 1) from the modified parallel analysis that the NOUN physics test goes beyond assessing test-takers proficiency in the course alone. Furthermore, a parallel analysis plot (See Figure 1) depicts the observed findings in the investigation [26].

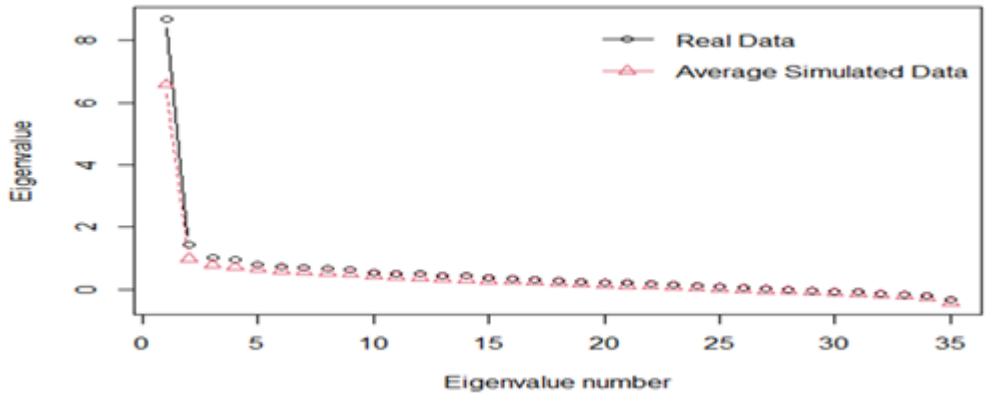

Figure 1: Parallel Analysis Plot for NOUN Physics Test Items

Figure 1 presented a schematic representation of eigenvalues from the observed and simulated NOUN physics test items dataset. The plot remarked a difference between the second eigenvalues of the observed data and the assumed IRT simulated dataset. This signifies that the elbow in the second eigenvalues should be considered strong evidence of more than one latent ability since it is greater than the eigenvalue of the simulated dataset. Thus, it concluded that PHY101 items are sufficiently multidimensional. That, factors other than physics ability accounted for test-takers responses to the Physics test.

Also, conditional independence of NOUN physics test items was assessed to see if items give clues to another item in the same test when ability (Θ) is conditioned. The conditional independence test was achieved using Q3 statistics implemented with the R language's multidimensional item response theory (mirt) package since the dataset shows evidence of multidimensionality [27]. The ltm package assumed a unidimensional dataset, so establishing conditional independence is impossible. Q3 is estimated for each item pair using the “residuals ()” function in R. The output is a 35 by 35 symmetric matrix (because the data set has 35 items). Q3 is a correlation of residuals; it ranges from –1 to 1. A significant absolute value indicates a more severe degree of local dependence. An abridged result of Q3 statistics is presented in Table 3.

Table 3: Abridge Q3 statistics of NOUN physics test items

Item	var1	var2	var3	var4	var5	var6	var7	var8	var9	var10	+	var31	var32	var33	var34	var35
var1	NA										+					
var2	0.05	NA									+					
var3	0.08	-0.06	NA								+					
var4	0.02	-0.05	0.00	NA							+					
var5	-0.03	-0.04	0.03	0.03	NA						+					
var6	-0.05	-0.05	0.03	-0.02	-0.01	NA					+					
var7	-0.09	0.01	0.05	0.02	-0.01	0.01	NA				+					
var8	0.02	-0.06	-0.02	-0.01	-0.01	0.02	0.02	NA			+					
var9	0.04	-0.01	-0.04	-0.05	0.18	-0.10	-0.08	0.04	NA		+					
var10	-0.03	-0.07	0.01	0.04	-0.02	-0.09	0.02	0.15	-0.03	NA	+					
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
var31	-0.01	-0.03	0.02	-0.03	0.04	-0.01	0.24	-0.05	-0.02	0.02	+	NA				
var32	-0.01	-0.01	0.34	0.25	0.03	-0.02	0.01	-0.01	0.04	0.02	+	0.31	NA			
var33	0.03	0.13	0.11	0.03	0.02	0.00	0.02	-0.01	-0.04	0.01	+	-0.16	0.10	NA		
var34	0.05	-0.05	0.18	0.20	-0.03	-0.03	0.24	-0.02	0.06	0.05	+	0.09	0.16	0.04	NA	
var35	0.13	0.20	0.01	-0.02	-0.05	0.01	0.12	0.01	-0.06	-0.02	+	0.11	0.09	-0.04	0.09	NA

Table 3 presents the Q3 statistics for physics test items. Yen (1993) recommended the absolute value of Q3 greater than 0.20 as a rule of thumb for flagging the local dependence of item pair. Examination of Table 3 yielded 595 residual correlations of item pairs, out of which ten-item pairs (for example, 3, 32; 4, 32; 7, 31; 7, 34; 31, 32, etc.) had Q3 above the cut-off. This implies that these items are performing the same function in the test, thus requiring one of them to be deleted. Consequently, it was concluded that few of the physics test items violated the assumption of conditional independence of the IRT framework [28].

Also, the monotonicity assumption of IRT was assessed using the function plot ( ) in R. Figure 2 presented the item characteristics curve for all the physics items.

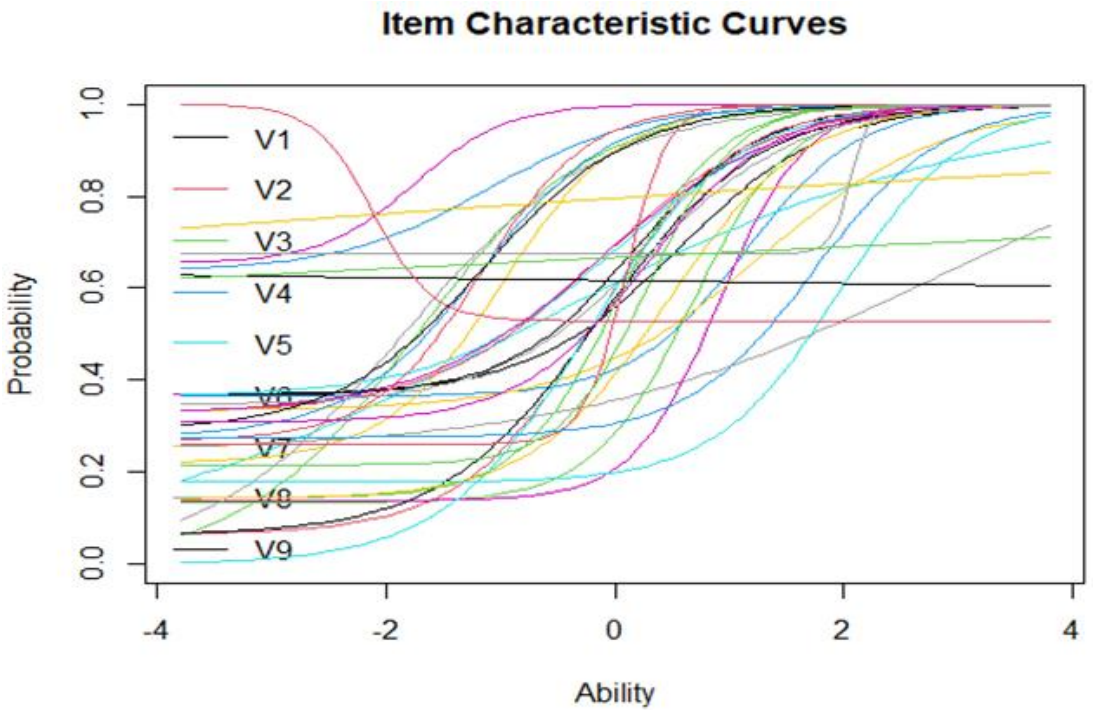

Figure 2: ICC plot for NOUN Physics Test Items

Figure 3 showed the relationship between the test-takers likelihood of responding to each item correctly and their underlying latent variable. It remarked that the more accessible item's functions are on the left side of the plot, in the lower regions of the latent trait scale, while the more complicated items are on the right side of the scale. As the test-takers ability increases on the continuum, their probability of getting the item correctly increases as well. The θ value at the inflection of each item's curve (at θ = 0.50) is the item's difficulty parameter.

7. DISCUSSION

Within the purview of the IRT framework, assumptions of the theory are fundamental to establish a better pathway to follow for the rest of the study. This paper assessed the underlying assumptions of IRT, such as dimensionality, conditional independence, and monotonicity, respectively. First, the dimensionality of NOUN physics test items was evaluated using a package in R language open-source software. When the analysis of the NOUN physics test data was begun, the anticipated outcome was that there would be one predominant dimension. The test would be classified as measuring nothing but test-takers proficiency in PHY 101. Surprisingly, the modified parallel analysis remarked that the test inherently had two or more relatively distinct ability dimensions. The items on the test could be classified into content categories based on which dimensions were required for successful performance. In addition, to further confirm the earlier observed dimensions from modified parallel analysis, one and two-dimensions were hypothesised and compared by the log-likelihood ratio test and information criteria. The p-value of the log-likelihood ratio test is virtually zero, rejecting the reduced, one-dimensional model. The descriptive model fit indices, AIC and BIC, in the two-dimensional model are lesser than those from the one-dimensional model. The results indicated that a two-dimensional solution yielded a better fit than a one-dimensional solution. Thus, test data was concluded to be multidimensional.

This indicates that more than one parameter (theta) reflects for test-takers location on a continuous latent variable. This is realistic to conclude that test-takers responses to the test items are due to their locations on multiple latent variables. There are various ways by which multidimensionality can be evident in the test. For instance, the physics 101 instrument used was designed to assess test-takers ability in physics. This test may have dimensions such as physics proficiency, reading proficiency, anxiety, interest, items from various themes, etc., which might be responsible for their observed responses to the test items. Also, test-takers with highly inclined reading ability might compensate to some extent for their lower physics ability to correctly respond to the test item. Thus, physics items indicate a compensatory multidimensional situation where test-takers location(s) on one or more dimensions can compensate for their locations on their latent variable(s). Findings from this study give

credence to the work of Emons, Sijtsma, and Pedersen (2012); Li, Jiao and Lissitz (2012); Bastug (2016); Oguoma, Metibemu, and Okoye (2016); Author that many different skills and knowledge are required to identify the correct answer in multiple-choice test items. The score for the test item is aims to describe individual interacting with the item, whether he/she has adequate skills and knowledge to select the correct answer, or that person is deficient in some critical component. That vital component could be reading skills, vocabulary knowledge, or the testing process using multiple-choice items. However, the study is against the submission of researchers such as Finch and Monahan (2008); Zhang and Stout (2016); Sijtsma and van der Ark (2017); Okwilagwe and Ogunrinde (2017); Author; Author that unidimensional factor explaining the variance observed in the performance of students test scores [27].

Another assumption assessed is conditional independence using Q3 statistics. This is the descriptive index of the conditional independence violation recommended by (Yen 1984). It correlates the residuals from a pair of items based on the difference between the observed response and the model-predicted response. According to De Ayala (2009), the behavior of Q3 is influenced by the number of items and the sample size; there is no single uniformly accepted cut-off value of Q3 [29]. A resampling technique such as a parametric bootstrap was used. It appeared to be a better approach to decide the cut-off of Q3 for a given data set (NOUN physics test items) under analysis (De Ayala, 2009). The findings showed that few test items violated the suggested absolute benchmark of Q3 greater than 0.20 as a rule of thumb for flagging conditional dependence of item pair. These items perform the same role in the test, thereby, one of them needs to be deleted or altered, and possible causes for the model violation should be assessed [30].

CONCLUSION

Test items are often derived from the construct's definition intended to be measured, in this case, physics proficiency. This item development approach ensures that all items capture the construct aimed at and only this construct. Thus, assessment of dimensionality and conditional independence is necessary for gathering evidence to support the validity of interpretations of scores from the test. This study conducted dimensionality and conditional independence of NOUN undergraduate physics test items based on this assertion. It concluded that the test possessed more than one trait to interpret students' test performance (more than one ability is measured by the items that make up the physics test). Also, a few pair of items show the dependency (that is, responses to the pair of items are statistically dependent). In other words, this means that the abilities specified in the model are not the only factors influencing students' responses to test items. This study recommends that the DEA embrace best practices to develop its test items. Training should be organised for lecturers on developing valid items, psychometric units should be created in the DEA of the university, and test and measurement experts should be employed. The major limitation is that only one course was examined despite numerous courses done at the undergraduate level. Further research could assess the assumptions of IRT on other courses.

Acknowledgment

The authors are very grateful to the management of the National Open University of Nigeria and the directorate of examination and assessment for the prompt release of students' responses used in this study.

Funding

This research received external funding provided by the Senate Research Grant awarded by the National Open University of Nigeria, Ref: NOUN/DRA/LARTL/005/VOL1.

Data Availability Statement

The dataset presented in this study are available on request. The data are not publicly available due to privacy reasons.

Conflicts of Interest

The authors declare no conflict of interest.

REFERENCES

[1] J. O. Amusa, M. A. Ayanwale, A. Ibrahim Oladejo, and F. Ayedun, "Undergraduate Physics Test Dimensionality and Conditional Independence: Perspective from Latent Traits Model Package of R Language," *Int. J. Assess. Eval.*, vol. 29, no. 2, pp. 47–61, 2022, doi: 10.18848/2327-7920/CGP/v29i02/47-61.

[2] M. A. Ayanwale, J. Chere-Masopha, and M. C. Morena, "The Classical Test or Item Response Measurement Theory: The Status of the Framework at the Examination Council of Lesotho," *Int. J. Learn. Teach. Educ. Res.*, vol. 21, no. 8, pp. 384–406, Aug. 2022, doi: 10.26803/ijlter.21.8.22.

[3] B. K. OLADELE, "COMPARISON OF SECONDARY SCHOOL STUDENTS'ABILITY IN MATHEMATICS CONSTRUCTED-RESPONSE ITEMS UNDER CLASSICAL TEST AND ITEM RESPONSE MEASUREMENT THEORIES IN THE IBADAN METROPOLIS, NIGERIA." 2021.

[4] A. L. Dima, "Scale validation in applied health research: tutorial for a 6-step R-based psychometrics protocol," *Heal. Psychol. Behav. Med.*, vol. 6, no. 1, pp. 136–161, Jan. 2018, doi: 10.1080/21642850.2018.1472602.

[5] M. Brucato, A. Frick, S. Pichelmann, A. Nazareth, and N. S. Newcombe, "Measuring Spatial Perspective Taking: Analysis of Four Measures Using Item Response Theory," *Top. Cogn. Sci.*, vol. 15, no. 1, pp. 46–74, Jan. 2023, doi: 10.1111/tops.12597.

[6] M. Robinson, A. M. Johnson, D. M. Walton, and J. C. MacDermid, "A comparison of the polytomous Rasch analysis output of RUMM2030 and R (ltm/eRm/TAM/lordif)," *BMC Med. Res. Methodol.*, vol. 19, pp. 1–12, 2019.

[7] S. Zegota, T. Becker, Y. Hagmayer, and T. Raupach, "Using item response theory to appraise key feature examinations for clinical reasoning," *Med. Teach.*, vol. 44, no. 11, pp. 1253–1259, Nov. 2022, doi: 10.1080/0142159X.2022.2077716.

[8] J. Brzezińska, "Item response theory models in the measurement theory," *Commun. Stat. Comput.*, vol. 49, no. 12, pp. 3299–3313, 2020, doi: https://doi.org/10.1080/03610918.2018.1546399https://doi.org/10.1080/03610918.2018.1546399.

[9] F. Carlberg Rindestig, M. Wiberg, J. E. Chaplin, E. Henje, and I. Dennhag, "Psychometrics of three Swedish physical pediatric item banks from the Patient-Reported Outcomes Measurement Information System (PROMIS)®: pain interference, fatigue, and physical activity," *J. Patient-Reported Outcomes*, vol. 5, no. 1, p. 105, Dec. 2021, doi: 10.1186/s41687-021-00382-2.

[10] J. W. B. Lang and L. Tay, "The science and practice of item response theory in organizations," *Annu. Rev. Organ. Psychol. Organ. Behav.*, vol. 8, pp. 311–338, 2021.

[11] E. N. Aiello *et al.*, "The Montreal Cognitive Assessment (MoCA): updated norms and psychometric insights into adaptive testing from healthy individuals in Northern Italy," *Aging Clin. Exp. Res.*, vol. 34, no. 2, pp. 375–382, Jul. 2021, doi: 10.1007/s40520-021-01943-7.

[12] A. Acevedo-Mesa, J. N. Tendeiro, A. Roest, J. G. M. Rosmalen, and R. Monden, "Improving the Measurement of Functional Somatic Symptoms With Item Response Theory," *Assessment*, vol. 28, no. 8, pp. 1960–1970, Dec. 2021, doi: 10.1177/1073191120947153.

[13] M. Johansson, M. Preuter, S. Karlsson, M.-L. Möllerberg, H. Svensson, and J. Melin, "Valid and reliable? Basic and Expanded Recommendations for Psychometric Reporting and Quality Assessment," *OSF Prepr.*, 2023.

[14] P.-C. Bürkner, "Bayesian item response modeling in R with brms and Stan," *arXiv Prepr. arXiv1905.09501*, 2019, doi: https://doi.org/10.48550/arXiv.1905.09501.

[15] S. Lipovetsky, "Handbook of Item Response Theory: Applications (3)," *Technometrics*, vol. 63, no. 3, pp. 433–437, Jul. 2021, doi: 10.1080/00401706.2021.1945327.

[16] L. E. de Ruiter and M. U. Bers, "The Coding Stages Assessment: development and validation of an instrument for assessing young children's proficiency in the ScratchJr programming language," *Comput. Sci. Educ.*, vol. 32, no. 4, pp. 388–417, Oct. 2022, doi: 10.1080/08993408.2021.1956216.

[17] J. Everaert, J. N. Vrijsen, R. Martin-Willett, L. van de Kraats, and J. Joormann, "A meta-analytic review of the relationship between explicit memory bias and depression: Depression features an explicit memory bias that persists beyond a depressive episode.," *Psychol. Bull.*, vol. 148, no. 5–6, pp. 435–463, May 2022, doi: 10.1037/bul0000367.

[18] J. A. Teresi *et al.*, "Evaluation of the measurement properties of the Perceived Stress Scale (PSS) in Hispanic caregivers to patients with Alzheimer's disease and related disorders," *Int. psychogeriatrics*, vol. 32, no. 9, pp. 1073–1084, 2020.

[19] J. E. Black, "An IRT Analysis of the Reading the Mind in the Eyes Test," *J. Pers. Assess.*, vol. 101, no. 4, pp. 425–433, Jul. 2019, doi: 10.1080/00223891.2018.1447946.

[20] S. Ghahraki, M. Tavakoli, and S. Ketabi, "Applying a two-parameter item response model to explore the psychometric properties: The case of the ministry of Science, Research and Technology (MSRT) high-stakes English Language Proficiency test," *Two Q. J. English Lang. Teach. Learn. Univ. Tabriz*, vol. 14, no. 29, pp. 1–26, 2022.

[21] S. Muranaka, H. Fujino, and O. Imura, "Evaluating the psychometric properties of the fatigue severity scale using item response theory," *BMC Psychol.*, vol. 11, no. 1, p. 155, May 2023, doi: 10.1186/s40359-023-01198-z.

[22] C. C. von Stülpnagel *et al.*, "Assessing the quality of life of people with chronic wounds by using the cross-culturally valid and revised <scp>Wound-QoL</scp> questionnaire," *Wound Repair Regen.*, vol. 29, no. 3, pp. 452–459, May 2021, doi: 10.1111/wrr.12901.

[23] F. Vuillermin and S. Huck-Sandhu, "Strategic planning in dynamic environments: how design thinking can complement corporate communication," *J. Des. Think.*, vol. 2, no. 2, pp. 85–96, 2021, doi: https://doi.org/10.22059/jdt.2021.323220.1056.

[24] A. Mwesiga and E. O. Okendo, "Effectiveness of Heads of Schools in Supervising Teachers' Teaching Activities in Secondary Schools in Kagera Region, Tanzania," *Int. J. Sci. Res. Manag.*, vol. 6, no. 04, p. 3390, Apr. 2018, doi: 10.18535/ijsrm/v6i4.sh04.

[25] D. Kartini, M. Kristiawan, H. Fitria, S. Negeri, and M. Sugihan, "The influence of principal's leadership, academic supervision, and professional competence toward teachers' performance," *Int. J. Progress. Sci. Technol.*, vol. 20, no. 1, pp. 156–164, 2020.

[26] A. Purwanto, J. T. Purba, R. Sijabat, and I. Bernarto, "The role of transformational leadership, organizational citizenship behaviour, innovative work behaviour, quality work life, digital transformation and leader member exchange on universities performance," *Linguist. Antverp.*, 2021, [Online]. Available: https://ssrn.com/abstract=3987666.

[27] S. Y. Pratama, J. Nurkamto, and A. Wijayanto, "The Representation of Multicultural Values in National Mandatory English Textbooks Used in Indonesian Secondary Schools," *Int. J. Multicult. Multireligious Underst.*, vol. 8, no. 1, p. 472, Jan. 2021, doi: 10.18415/ijmmu.v8i1.2337.

[28] F. Firdaus, D. K. Anggreta, and F. Yasin, "Internalizing Multiculturalism Values Through Education: Anticipatory Strategies for Multicultural

Problems and Intolerance in Indonesia," *J. Antropol. Isu-Isu Sos. Budaya*, vol. 22, no. 1, p. 131, May 2020, doi: 10.25077/jantro.v22.n1.p131-141.2020.

[29] A. Halik, S. W. Hanafie Das, M. S. Dangnga, M. Rady, M. Aswad, and M. Nasir, "Empowerment of School Committee in Improving Education Service Quality at Public Primary School in Parepare City," *Univers. J. Educ. Res.*, vol. 7, no. 9, pp. 1956–1963, 2019.

[30] J. L. Mahoney *et al.*, "Systemic social and emotional learning: Promoting educational success for all preschool to high school students.," *Am. Psychol.*, vol. 76, no. 7, pp. 1128–1142, Oct. 2021, doi: 10.1037/amp0000701.

DOI: https://doi.org/10.15379/ijmst.v10i5.2531