Interchange (2024) 55:137–157
https://doi.org/10.1007/s10780-024-09517-y

Multidimensional item Response Theory Calibration of Dichotomous Response Structure Using R Language for Statistical Computing

Musa Adekunle Ayanwale[1] · Jamiu Oluwadamilare Amusa[2] · Adekunle Ibrahim Oladejo[3] · Funmilayo Ayedun[2]

Received: 17 October 2022 / Accepted: 20 February 2024 / Published online: 6 March 2024

Abstract

The study focuses on assessing the proficiency levels of higher education students, specifically the physics achievement test (PHY 101) at the National Open University of Nigeria (NOUN). This test, like others, evaluates various aspects of knowledge and skills simultaneously. However, relying on traditional models for such tests can result in inaccurate interpretations of students' abilities. The research highlights the importance of exploring the multidimensional nature of the PHY 101 test to improve its accuracy in measuring student proficiency and enhance education and assessment quality at NOUN. Using an ex-post facto research design, the study analyzed 978 responses from NOUN's Directorate of Examination and Assessment. Through confirmatory and exploratory DETECT techniques, the study found strong evidence supporting the test's multidimensionality. Three distinct dimensions emerged: cognitive processing, reading ability, and problem-solving skills. A parsimonious multidimensional three-parameter logistic model was used to calibrate the test items, providing valuable insights into item difficulty, discrimination, and resistance to chance influences. While the study primarily focuses on the psychometric aspects of the PHY 101 test, it is important to consider its broader impact on the educational community. The research contributes to educational assessment by emphasizing the significance of recognizing and addressing the multidimensional nature of higher education tests. This approach can result in more accurate assessments of students' abilities, ultimately improving education quality and fairness. The findings confirm the multidimensional nature of the PHY 101 test and identify three distinct dimensions, aligning with the study's objective. These insights are relevant to educators and test developers, highlighting the need for a multidimensional approach to effectively assess and enhance student proficiency. For researchers interested in similar studies, it is recommended to explore the broader influence of multidimensional models in educational assessment. Investigating their impact on teaching methods, curriculum development, and student learning experiences can provide valuable

Extended author information available on the last page of the article

insights. Longitudinal studies assessing the long-term effects of multidimensional assessment on student outcomes and success are also recommended.

Keywords Multidimensional item response theory · Multidimensional three parameters logistic · Dimensionality physics achievement test · mirt Package · R Language

Introduction

The use of exams in higher education institutions (HEIs) to gather reliable and meaningful information has become increasingly important. Multiple-choice questions are commonly used to assess students' performance, including in undergraduate exams at Nigeria's National Open University (NOUN), especially in subjects like PHY 101, a course in physical science. This assessment method is widely favored in the educational community due to its efficiency, reliability, and ease of scoring. The quality of these multiple-choice questions is crucial as they directly impact students' assessment outcomes, reflecting their overall competency level (Amusa et al., 2022; Ayanwale et al., 2020; Ayanwale & Adeleke, 2020). In terms of Bloom's taxonomy, high-quality multiple-choice questions can assess advanced cognitive skills such as interpretation, critical thinking, application, and synthesis (Akinboboye & Ayanwale, 2021; Owolabi et al., 2023). However, crafting such questions can be challenging and requires careful construction to ensure their meaningfulness. Developing a comprehensive question bank through meticulous item analysis is considered a valuable resource for universities to conduct assessments. Undoubtedly, a test's psychometric properties provide valuable insights into its appropriateness, utility, and validity, ultimately determining its legitimacy (Ajeigbe & Afolabi, 2014; Ojerinde et al., 2012, as cited in Adekunle et al., 2021; Ayanwale et al., 2019). The complexities associated with obtaining valid and reliable psychometric properties for these test items highlight the importance of employing sophisticated and precise analytical methods. At the beginning of test development, item response theory (IRT) is used to explore test dimensionality and establish the validity foundations that support the test's purpose, usage, and inferences about test-takers (Ayanwale, 2021; Ayanwale & Ndlovu, 2021; Amusa et al., 2022). Essentially, test dimensions correspond to the latent traits that developers aim to measure. Items are carefully created and organized to align with these intended traits or dimensions. This rationale supports the adoption of multidimensional item response theory (MIRT) to calibrate the NOUN physics test. Using IRT, a test-taker's response to a specific question depends on an unobservable trait or ability within their mind. It is assumed that various latent traits or abilities exist along a continuous dimension, ranging from the lowest to the highest (referred to as θ). The test-taker's position on this dimension, represented as θi, is commonly known as their ability or proficiency. According to IRT, as the number of items increases, one can expect a monotonous increase in the probability of correctly answering a question. IRT models are particularly applicable to binary-scored items where responses are categorized as either correct or incorrect. These models

are generally known as unidimensional models and are suitable when all test items aim to measure the same underlying capability dimension. Unidimensional IRT (UIRT) is based on the assumption that each test item measures a latent trait, also known as a common underlying ability. When using overall test scores as an assessment criterion for different ability levels, it is important that a test designed to measure one trait is not influenced by other traits. According to Hattie (1985) cited in Ackerman et al. (2003) and Ayanwale et al. (2022), no technique can provide satisfactory results for unidimensionality under different conditions. Monte Carlo studies have also cautioned against interpreting results from dimensionality indexes in other contexts (test length, sample size, etc.; De Champlain & Gessaroli, 1998; Gessaroli & De Champlain, 1996; Hattie et al., 1996 cited in Sheng & Wikle, 2007). Research has shown that parameter estimation becomes biased when a multidimensional test is modeled using a unidimensional model (Immekus et al., 2019; Wiberg, 2012). Additionally, measurement errors can increase, making it difficult to accurately infer a student's proficiency in a given subject (Walker & Beretvas, 2000 cited in Sheng & Wikle, 2007). Consequently, interpretations of test scores become questionable as they are not considered useful, meaningful, or appropriate.

Multidimensional item response theory (MIRT) models are used when a manifest response for an item is influenced by multiple abilities (more than one Θ). Tests in fields such as statistics, physics, education, or psychology may have multiple dimensions or constructs. A construct is a theoretical representation of a dimension and is commonly modeled using MIRT (Zhang & Stone, 2008). MIRT models are appropriate because they predict an examinee's likelihood of answering a specific question by considering latent (unobserved) variables. According to Reckase (1997, 2009), MIRT models are popular tools for assessing test content, item calibrations, and computerized adaptive tests. Similar to unidimensional models, MIRT models have certain assumptions, including monotonicity and local independence. Monotonicity implies that the probability of answering an item correctly increases as the student's ability level increases (Smith, 2009 cited in Kose & Demirtasli, 2012; Ul Hassan & Miller, 2022). Under the local independence assumption, the probability of an item's response is independent of other item responses, regardless of item and person parameters. In the field of Multidimensional Item Response Theory (MIRT), models can be classified as compensatory or non-compensatory. This classification depends on whether a high level of proficiency in one trait can make up for a lower level in another (Sijtsma & Junker, 2006; Kose & Demirtasli, 2012). Compensatory MIRT models are a sophisticated approach used in educational and psychological assessments. They are designed to address situations where test items are influenced by multiple underlying abilities or traits (Sheng & Wikle, 2007). These models are called "compensatory" because they take into account the idea that a deficiency in one ability can be balanced or compensated for by a higher level of proficiency in another. In practical terms, this means that compensatory MIRT models recognize that many test questions may require a combination of skills or knowledge areas to provide correct answers. For example, consider a physics test question that requires both mathematical reasoning and knowledge of physics. In such cases, a student can compensate for a lack of strong physics knowledge by possessing exceptional math skills. Within compensatory MIRT models, each test item has two fundamental

parameters: difficulty and discrimination. The difficulty parameter represents the proficiency level at which an individual has a 50% probability of answering the item correctly. Conversely, the discrimination parameter evaluates how effectively the item distinguishes between individuals with different levels of the latent trait or ability. These models estimate these parameters for each underlying ability. Scoring in compensatory MIRT typically involves combining scores from various latent traits. The final score provides an overall assessment of an individual's performance, taking into account the compensatory nature of the model. This holistic approach enables a more comprehensive evaluation of a test-taker's abilities. Compensatory MIRT models are widely used in educational assessments, especially when it is evident that answering a single test item correctly depends on a combination of abilities (Bolt & Lall, 2003; Immekus et al., 2019). They excel in capturing the intricate interplay between diverse skills and knowledge domains. However, it is important to acknowledge that these models can be computationally demanding, especially when dealing with a large number of latent traits or complex item response data. Additionally, interpreting outcomes from compensatory models may be less straightforward compared to non-compensatory models, due to their inclusion of interactions between latent traits. In summary, compensatory MIRT models serve as invaluable tools for evaluating individuals in situations where abilities interact and compensate for each other. They offer a more authentic and precise portrayal of a test-taker's abilities (Reckase, 2009). This makes them a valuable asset in the field of educational and psychological measurement.

Conversely, non-compensatory Multidimensional Item Response Theory (MIRT) models are a significant aspect of educational and psychological assessments. These models are designed to address situations where test items do not allow one ability to make up for a deficiency in another. In other words, in non-compensatory MIRT models, proficiency in one ability cannot fully compensate for a deficit in another (Embretson & Reise, 2000). Consider a test item that requires strong reading comprehension and physics knowledge. In non-compensatory MIRT models, excelling in one area cannot compensate for a lack of competence in the other. Therefore, achieving a high score on such an item requires proficiency in both reading and physics. Within non-compensatory MIRT models, test items are characterized by their difficulty and discrimination parameters for each underlying ability. The difficulty parameter indicates the proficiency level at which there is a 50% chance of correctly answering the item. The discrimination parameter reflects the item's ability to differentiate between individuals with varying levels of the latent trait or ability. Scoring in non-compensatory MIRT typically involves assessing each latent trait separately. Instead of combining scores from different abilities, the focus is on evaluating each ability independently. This approach can provide a more detailed view of a test-taker's strengths and weaknesses in each dimension. Non-compensatory MIRT models are suitable when it is crucial to maintain a clear distinction between distinct abilities. For instance, if a test aims to assess both reading and physics knowledge separately without allowing one skill to compensate for the other, non-compensatory models are preferred. Such models are often used when the relationships between abilities are well-defined and should not be blurred. However, it's important to recognize that implementing non-compensatory MIRT models can be challenging due

to the absence of efficient algorithms for estimating item parameters when abilities are interrelated. Additionally, the interpretation of results from non-compensatory models is more straightforward but may provide a less holistic view of a test-taker's abilities compared to compensatory models. Importantly, compensatory models are more prevalent in educational research, even though both compensation and non-compensation models are used (Drasgow & Parsons, 1983; Kose & Demirtasli, 2012; Ozdemir & Gelbal, 2022; Robitzsch, 2020). This preference might be because non-compensatory models lack efficient algorithms for estimating item parameters. In the context of Multidimensional IRT, the Item Characteristic Surface (ICS) illustrates the probability that a test-taker will correctly answer an item based on their composite ability. Recent applications of MIRT often use fewer dimensions, typically two, due to limitations in estimation programs. When item responses depend on more than two latent traits, they are referred to as "item response hyper-surfaces." Visualizing ICS in multidimensional latent space becomes challenging when dealing with more than three dimensions. Despite the prevalence of MIRT, unidimensional Item Response Theory (UIRT) has been extensively used in education and psychology for decades. Many achievements tests struggle to meet the assumption of unidimensionality, leading to the continued use of UIRT (Ackerman, 2010; Algina & Swaminathan, 2015; Hambleton & Swaminathan, 1985; Kose & Demirtasli, 2012; Ul Hassan & Miller, 2022; Yang, 2007).

The comparison of unidimensional and multidimensional models has been conducted by many researchers (Spencer, 2004; De La Torre & Patz, 2005; Yang, 2007; Kose & Demirtasli, 2012; Reckase, 1985; Sympson, 1978; Mulaik, 1972; Hamsy, 2014; Ha, 2017; Zulaeha et al., 2020; Zhang, 2004; Liu et al., 2013). As a result of the increased number of latent traits that influence item performance, item parameters have provided more accurate measurements under MIRT. Moreover, a comparison between MIRT and data-driven models shows that model-data fit indexes favor MIRT models. An examination of the UIRT models based on multidimensional college admission test data was undertaken by Wiberg (2012). A simulation study showed that MIRT gives better results than UIRT in modeling fit. However, UIRT is similar to MIRT when conducted consecutively. According to her, if there were multidimensionality between items in the test, consecutive UIRT models instead of MIRT models might be better suited and easier to interpret. A paradox arising with MIRT compared to UIRT has been highlighted by Hooker et al. (2009): if an examinee changes his or her answer to an item from correct to incorrect, it could result in a decrease in the estimated parameter (Finkelman et al., 2010; Hooker, 2010; Jordan & Spiess, 2012, 2018). Additionally, Kose and Demirtasli (2012) found that MIRT is more accurate than UIRT at estimating these latent traits since the standard error of MIRT is smaller than that of UIRT. A test's standard error of model parameters decreases the more items it has, which is an important consideration for educators when designing tests, according to Kose and Demirtasli. Li et al. (2012) used the Multidimensional 2-Parameter Logistic (M2PL) approach for dimensionality validation of a K-12 science assessment. In addition to the unidimensional IRT and testlet models, practitioners preferred multiple-dimensional estimates. Another example illustrates the preference for a multidimensional model over a unidimensional one. It is impossible to know in advance how the ability dimensions will work in

many test situations. As a result, choosing a model can be difficult. Consider a test in which the three components are listening, reading, and writing. There appears to be a connection between the three subtests. The items can be thought of as measuring a single English ability dimension and fitting a unidimensional model. The estimates might be biased if the subtests are not measured exactly similarly. To estimate the examinee's ability in each sub-dimension, the unidimensional model may be fitted separately for each subtest. Due to the fact that the unidimensional model does not account for the relationships between different ability dimensions, these sub-scores apply only when there is no correlation between the subtests. The one-dimensional model restricts both approaches in this sense. A multidimensional model would provide more precise estimates and therefore be more efficient since it can draw strength from the responses of correlated ability dimensions (Finch, 2011; Liu et al., 2013; Sheng & Wikle, 2007; Zhang, 2004; Zulaeha et al., 2020). Meanwhile, since model-data fit assessment conducted supported M3PLM as the most parsimonious, the next paragraph discusses the model and parameter estimation in detail.

Multidimensional 3-parameter Logistic Model

Unlike unidimensional IRT, multidimensional IRT allows simultaneous analysis of multiple constructs. Multiple ability dimensions model the probability of success in MIRT. A vector $\theta_{j=(\theta_{j1},\ldots\ldots\ldots\ldots,\theta_{jk})}$ represents the ability parameter values for each individual. In Reckase (2009), items are classified based on discrimination parameter values $a_{i1=(a_{i1},\ldots\ldots\ldots\ldots,\theta_{ik})}$, difficulty parameter $d_i = -a_i b_i$, and ci, a lower asymptote parameter. M3PL was, therefore, implemented using this mathematical expression.

$$P\left(X_{ij} = 1/\theta_j, a_i, c_{i,} d_i\right) = C_i + \left(1 - C_i\right)\frac{exp\left(a_i.\widehat{\theta}_j + d_i\right)}{1 + exp\left(a_i.\widehat{\theta}_j + d_i\right)} \tag{1}$$

where exp(.) is the exponential function with base e.

$P\left(X_{ij} = 1/\theta_j, a_i, c_{i,} d_i\right)$ is the probability of student j's correct response to item i.

θ_j is vector of student j's ability.

a_i is vector of item i slope.

$c_i \in (0, 1)$is guessing parameter.

$d_i$is the intercept parameter, and vectors $a_i.\widehat{\theta}_j$have the same elements m, which is the number of dimensions.

The M3PL model was developed to account for empirical findings, such as those reported by Lord (1980), which indicate that examinees with low abilities are less likely to answer multiple-choice questions correctly. In this model, a single lower asymptote parameter is used to specify the probability of a correct response for examinees with very low values of θ. The process of selecting a correct response for individuals with low capabilities does not seem to be related to the constructs assessed by the test item. The interval of ci theoretically ranges between 0 and 1. However, in practice, Baker

(2001), Baker and Kim (2017), and Seock-Ho (2004) suggest that ci ranges from 0 to 0.35 as an acceptable item calibration cut-off.

MIRT has been updated with some new concepts, such as the item characteristics surface (ICS). Reckase (1985, 1997) describes the concepts of multidimensional item difficulty (MDIFF) and multidimensional item discrimination (MDISC) as being incompatible with each other. MDISC is an overall measure that quantifies the level of discrimination on multiple dimensions and is analogous to the 'a' parameter in UIRT. In the ICS, a longer vector indicates a more discriminating item. In the model, the discrimination parameter measures how different items are. It is considered valuable if an item discriminates well between subjects with different abilities and interests in the exam. As ability increases, the probability of correctly answering a question increases with a higher discrimination parameter value. An ICS will be steeper with a high discrimination parameter value, allowing the item to differentiate subjects better around its difficulty level. If the item difficulty is within the scope of the exam, items with a high discrimination power can contribute more to assessment precision than items with a low discrimination power (Ackerman, 1996; Ackerman et al., 2003; Ha, 2017; Kose & Demirtasli, 2012; Ul Hassan & Miller, 2022). Therefore, an item's ability to discriminate between examinees is important for educators. MIRT items differentiate examinees in each dimension, whereas UIRT items only discriminate in one direction. The discriminating power of item i for the most discriminating combinations of dimensions (M3PL) can be expressed as:

$$MDISC = \sqrt{\sum_{k=1}^{m} a^2{}_{ik}} \approx \sqrt{a^2{}_{i1} + a^2{}_{i2} + a^2{}_{i3}} \tag{2}$$

where $a1$, $a2$ *and* $a3$ are the discrimination parameter for each of item *i.*

More so, the MDIFF has the same meaning as *bi* in UIRT but is not interchangeable. Using IRTs, MDIFF is the distance from the origin to the vector's steepest slope (Smith, 2009; Ackerman et al., 2003; Ha, 2017; Reckase, 1985). Thus, the general expression for the distance of the line from the origin is given by:

$$MDIFF = \frac{-\left(d_i\right)}{\sqrt{\sum_{n=1}^{m} a_{in}^2}} \approx \frac{-(d_i)}{MDISC} \tag{3}$$

where *di* represents the intercept for item *i*. Depending on the sign of the distance, we can determine how difficult the item is. Positive MDIFF means that items are relatively hard in the first quadrant, while negative MDIFF means that items are relatively easy in the third quadrant (Ackerman et al., 2003; Ha, 2017; Mark et al., 1983; Reckase, 2007, 2009).

Current Study

This study aimed to investigate the dimensions of the physics test items and calibrate the test using both the parsimonious model based on Supplementary Item Response Theory (SIRT) and the Multidimensional Item Response Theory (MIRT) packages in the R programming language. High-quality test items are essential for assessing students' proficiency levels and overall competence. Previous research has shown that MIRT models offer promise and provide better parameter indices when items are associated with multiple latent traits. However, there has been limited exploration of parameter estimation when some items are linked to multiple dimensions, specifically using the M3PL model for calibration. The complexity of the physics test items at NOUN is a major concern since these items may not strictly adhere to a simple structure. Instead, they may exhibit various latent traits to different extents, regardless of their position in the test. Therefore, this study has four main objectives: (1) to determine the number of dimensions underlying the physics achievement test, (2) to understand the difficulty parameter of the test, (3) to estimate the discrimination parameter of the test, and (4) to assess the chance factor involved in the test.

Method

This study used a non-experimental ex-post facto design, chosen for its suitability in retrospectively describing an event and applying that description to verify its occurrence. In the context of Item Response Theory (IRT), the students' answers to the test items served as the response variables, while the examinees' abilities and the characteristics of the test items, as indicated by previous studies (Adekunle et al., 2021; Li et al., 2012), were considered latent predictor variables. Census sampling was employed to ensure a comprehensive representation of the target population, specifically, all undergraduate students enrolled in PHY101 across various study centers under the National Open University of Nigeria (NOUN). This approach was deemed appropriate as it covered the entire population under study.

The research instrument consisted of thirty-five items selected from the PHY 101 physics achievement test. These items were meticulously developed by faculty lecturers to assess students' comprehension of various physics concepts, including mechanics, thermodynamics, and electricity, among others. The instrument utilized a multiple-choice format, evaluating students' ability to analyze, reason, and apply physics concepts across a wide range of topics. Each item offered four answer options, with only one correct response allowed. Correctly answered items were coded as 1, while incorrectly answered items were coded as 0. Additionally, each test item measured three latent traits: reading ability, reasoning/cognitive-processing skill, and problem-solving in physics. These traits were taken into account during the testing process. It is worth emphasizing that the research instrument underwent a rigorous development process. Faculty lecturers responsible for item creation followed a systematic procedure to ensure the validity and reliability of the test items. This process included the review of content experts and alignment of items with the intended learning outcomes of PHY 101. Subsequently, pilot testing was conducted

to identify and address potential issues related to item clarity, difficulty, and relevance. Adjustments were made to the items based on the feedback received during pilot testing, ensuring their effectiveness in assessing the targeted knowledge and skills. Consequently, the instrument demonstrated empirical reliability with a coefficient of 0.80, and its content validity was established at 0.89. Data collection took place across NOUN study centers located in all 36 states of Nigeria. The Directorate of Examination and Assessment (DEA) at the University managed all aspects related to examinations and assessments. For data analysis, the study utilized examinee data from 978 responses, of which 37.8% were females and 62.2% were males. The study used a comprehensive approach to analyze the data and extract meaningful insights from secondary data obtained from the DEA. The primary analytical tools employed were the Supplementary Item Response Theory Models (sirt) package (Robitzsch, 2015, 2020) and the multidimensional item response theory (mirt) package (Chalmers, 2012) within the R language version 4.0.2. The function conf. detect was used to compute the DETECT under a confirmatory specification of item clusters (Stout et al., 1996; Zhang & Stout, 1999a, b). On the other hand, Exploratory DETECT was used when there were no predefined assumptions about the data's structure, aiming to explore potential dimensions or factors that might influence item responses.

Initially, the data was assessed for dimensionality to understand the underlying structure of the PHY 101 test. Techniques like factor analysis or principal component analysis were employed to reveal and isolate these hidden dimensions. The strength of the association between each item and these dimensions was examined, with strong associations indicating that the item measured that specific dimension. Labels and meanings were then assigned to these dimensions based on the content of items strongly associated with each factor. This process provided insights into the underlying structure of the data and shed light on distinct dimensions.

Confirmatory DETECT provided robust evidence supporting the test's multidimensional nature, while exploratory DETECT revealed three distinct dimensions: cognitive processing, reading ability, and problem-solving skills. After analyzing the dimensionality, the study focused on assessing the model-data fits to ensure the appropriateness of the chosen models. This step was critical in calibrating the multidimensional three-parameter logistic model for item parameters such as difficulty, discrimination, and susceptibility to chance influences. This comprehensive data analysis process contributed to the overall reliability and validity of the research findings by providing a deep understanding of the test's structure and the quality of its individual items.

Results

We examined the dimensionality of the items in NOUN PHY 101 by applying Stout's test of essential unidimensionality (Stout, 1987) to the responses of examinees on the test form. We used the RAM package Supplementary Item Response Theory Models (sirt) implemented in the R Language and Environment for Statistical Computing (R Core Team, 2019), as developed by Robitzsch (2015, 2020). Through both

confirmatory dimensionality evaluation (confirmatory DETECT) and exploratory dimensionality evaluation (exploratory DETECT), we determined the optimal number of dimensions to support the relevance of the test items based on the results from confirmatory dimensionality evaluation. DETECT, as used by Akerman (2003, p. 137), Ayanwale (2021, 2023), and Zhang (2013), is a nonparametric exploratory procedure for assessing dimensionality and estimating the magnitude of departure from unidimensionality. Consequently, Jang and Roussos (2007) and Zhang (2007) judge the dimensionality of test data using the classification indices below, while Table 1 presents the results of the test's dimensionality assessment.

Essential unidimensionality	DETECT < .20
Maximum value under simple structure	ASSI=1 RATIO=1
Essential deviation from unidimensionality	ASSI > .25 RATIO > .36
Essential unidimensionality	ASSI < .25 RATIO < .36.
Strong multidimensionality	DETECT > 1.00
Moderate multidimensionality	.40 < DETECT < 1.00
Weak multidimensionality	.20 < DETECT < .40

Table 1 illustrates the NOUN physics achievement test for undergraduates' dimensionality test. This table illustrates that the test items returned a DETECT value of 6.202, an ASSI value of 0.867, and a RATIO value of 0.863. There was strong evidence of multidimensionality in the NOUN physics achievement test, based on results obtained from undergraduate students (DETECT > 1.00). It is also of great significance to note that the outcome of this study showed the multidimensionality of the test to have a simple structure as the ASSI was approximately 1 (ASSI ≈ 1) and the RATIO was also approximately 1 (Ratio ≈ 1). Meanwhile, it is possible that DETECT might not produce optimal results in the absence of an approximate simple structure(Svetina & Levy, 2014). Using exploratory DETECT, further analysis of the students' responses to the physics test was conducted to understand the actual number of dimensions of the test. Figure 1 illustrates the result that was obtained.

Figure 1 illustrates the dimensions underlying the examinees' performance in the physics examination and the number of items loading under each dimension. The figure shows the three dimensions considered when estimating students' performance in the physics test. Based on the test's nature, researchers, and experts suggested that reading skills had a stronger impact on the first dimension of the test compared to cognitive processing/reasoning and problem-solving skills. In contrast, cognitive processing/reasoning ability was more dominant in the second dimension than reading ability and problem-solving skills, while problem-solving skill had a

Table 1 Dimensionality assessment of physics achievement test

	Unweighted	Weighted
DETECT	6.202	6.202
ASSI	0.867	0.867
RATIO	0.863	0.863

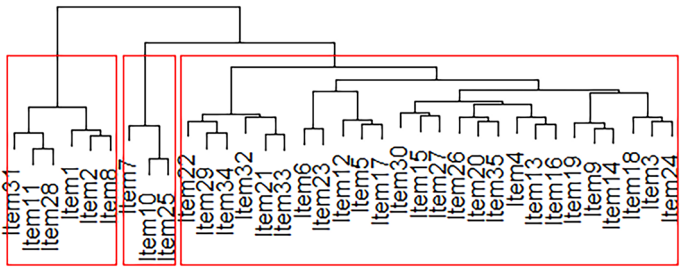

Fig. 1 Dimensionality of the NOUN physics achievement test

stronger impact on the third dimension of the test compared to cognitive processing/reasoning and reading skills. An assessment of model fit was also conducted on examinee responses to determine the most parsimonious model for estimating and calibrating item parameters for the physics test items. This was achieved using multidimensional item response theory (mirt) packages of full information factor analysis. These packages were used to hypothesize and compare the multidimensional 1-parameter logistic model (M1PL), multidimensional 2-parameter logistic model (M2PL), multidimensional 3-parameter logistic model (M3PL), and multidimensional 4-parameter logistic model (M4PL). The assessment of model-data fit for the physics achievement test is presented in Table 2.

A summary of the results presented by the full information factor analysis is presented in Table 2, which contains the indices of fit through the Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), Sample size Bayesian Information Criterion (SABIC), Hannan and Quinn's information criterion (HQ) and the −2loglikelihood ratio. As a result of the hypothesis comparison between the two models, the values for the indices of the M2PL solution (AIC = 32659.5, SABIC = 32893.7, HQ = 32914.2, BIC = 33328.9 and −2logLikelihood = −16,193) were less than those for the M1PL solution. Furthermore, when compared with M2PL solution, M3PL showed an improved fitness with (AIC = 32250.8, SABIC = 32544.9, HQ = 32570.5, BIC = 33091.1, and −2logLikelihood = −15,953), while M4PL shows an inferior fit when hypothesised with M3PL solution (AIC = 32236.5, SABIC = 32590.4, HQ = 32621.3, BIC = 33247.8, and −2log-Likelihood = − 15,911). The observed data does not match the predictions of the other models, which indicates a significant difference. Consequently, the M3PL IRT model was found to be the most appropriate one to calibrate NOUN physics achievement test. Next, we need to determine the level of difficulty of the NOUN physics

Table 2 Model-data fit assessment for physics achievement test

Model	AIC	SABIC	HQ	BIC	-2logLikelihood
M1PL	34600.5	34662.0	34667.4	34776.4	-17,264
M2PL	32659.5	32893.7	32914.2	33328.9	-16,193
M3PL	32250.8	32544.9	32570.5	33091.1	-15,953
M4PL	32236.5	32590.4	32621.3	33247.8	-15,911

achievement items. To do this, we calibrated examinee responses to the test using the most parsimonious M3PL IRT model that closely matched the test's characteristics. Firstly, it should be noted that MDIFF difficulty parameter represents the equivalent form of bi in the unidimensional IRT that were judged based on the principles laid out by Baker, 2001;, 2017; Hasmy, 2014 for judging the difficulty parameter of MDIFF. There are five levels of difficulty as defined by the authors. According to them, the item is extremely difficult when b or MDIFF ≥ 2, difficult when $0.5 \leq$ b or MDIFF < 2, moderately difficult when $-0.5 \leq$ b or MDIFF < 0.5, easy when $-2 \leq$ b or MDIFF < -0.5 and extremely easy when b or MDIFF < -2. In Table 3, we present the results of the study.

Based on the multidimensional calibration of the test, Table 1 shows that 13 (37.1%) of the items have a moderate difficulty index, while 2 (5.8%) of the items have an extremely difficult index, and 9 (25.7%) of the items have a difficulty index (hard items). Of the remaining 11 (31.4%) items, there are easy items. Considering the results of this study, a reasonable number of the items on the physics test were moderately difficult for the examinees to answer. Essentially, the results suggest that the items on the test are only appropriate for determining the proficiency of examinees who have a moderate level of proficiency in PHY 101 and those who are proficient in the course, depending on the test content. This test was not suitable for measuring the proficiency in physics of examinees who have a low level of knowledge about the course. The item ordering of the test has also been established in the literature as one of the factors that could affect the examinee's performance. As a rule, a good test should start with relatively easy items and progress to more complex ones, enabling examinees with low proficiency to start the exam with confidence instead of starting with moderately difficult items. Hence, it would appear that the pattern of difficulty of items revealed in Table 2 goes against the norms that should be followed when developing tests.

To evaluate the degree of discrimination that each item of the physics tests possesses across each dimension (a1, a2, and a3) in terms of discrimination against individuals, which is recognized as multidimensional item discrimination (MDISC), the discrimination parameter of the MDISC is the equivalent form of a.i. in the unidimensional IRT. The discrimination was evaluated following the criteria established by (Ayanwale et al., 2018; Ayanwale, 2019; Baker, 2001; Hasmy, 2014). To describe the item's discriminatory power, the authors have determined that it is very highly discriminatory if a or MDISC ≥ 1.7, highly discriminatory if $1.35 \leq$ a or MDISC < 1.7, moderately discriminating if $0.65 \leq$ a or MDISC < 1.34, lowly discriminating when $0.35 \leq$ a or MDISC < 0.65, and very lowly discriminating when a < -0.35. Detailed results are given in Table 4.

As shown in Tables 4 and 29 out of the 40 items (82.8%) have a very high discriminating index, 2 out of the 40 items (5.7%) have a high discriminating index, 3 out of 40 items (8.6%) have a moderately discriminating index, and we have one item (2.9%) with a low discriminating index out of the 40 items. There was also a positive discriminating index for all the test items, which indicates that as the examinee's abilities level increases, the probability that they will take correct responses increases. Due to the test's items, there was strong discrimination between examinees who were proficient in physics and those who were not proficient in the course.

Table 3 Three-multidimensional item difficulty parameter of physics tests

Item	d	a1	a2	a3	g	MIDFF	Remarks
Item1	− 0.87	0.64	1.01	2.64	0.12	0.30	Moderately difficult
Item2	0.49	− 0.38	− 1.73	0.87	0.01	− 0.25	Easy
Item3	− 6.35	− 2.78	− 0.64	− 1.14	0.18	2.07	Extremely difficult
Item4	0.36	− 2.96	− 0.95	− 0.34	0.20	− 0.12	Easy
Item5	0.88	− 4.01	− 0.59	− 0.31	0.26	− 0.22	Easy
Item6	− 0.17	− 1.26	− 1.35	− 0.60	0.26	0.09	Moderately difficult
Item7	0.47	0.25	0.91	1.66	0.14	− 0.25	Easy
Item8	0.49	− 0.09	− 1.77	1.18	0.12	− 0.23	Easy
Item9	− 0.04	− 1.06	− 2.52	1.40	0.12	0.01	Moderately difficult
Item10	− 0.20	− 1.34	− 0.66	2.21	0.00	0.07	Moderately difficult
Item11	− 0.43	− 1.50	0.04	1.81	0.02	0.18	Moderately difficult
Item12	0.61	− 0.48	− 1.97	0.48	0.15	− 0.29	Easy
Item13	− 0.77	− 0.76	− 1.43	− 0.21	0.20	0.47	Moderately difficult
Item14	− 0.59	− 0.75	− 1.73	− 0.12	0.40	0.31	Moderately difficult
Item15	0.66	− 0.84	− 1.75	− 2.05	0.22	− 0.23	Easy
Item16	0.07	0.15	0.44	0.02	0.35	− 0.15	Easy
Item17	− 1.31	− 1.55	− 1.30	0.19	0.34	0.64	Difficult
Item18	− 0.32	0.63	2.01	1.01	0.33	0.14	Moderately difficult
Item19	− 3.28	− 2.21	− 1.33	0.63	0.38	1.24	Difficult
Item20	− 1.08	0.64	− 2.72	− 0.66	0.29	0.38	Moderately difficult
Item21	− 1.20	1.02	− 1.16	− 1.81	0.37	0.50	Difficult
Item22	− 1.82	1.00	− 2.14	− 0.48	0.22	0.76	Difficult
Item23	− 3.82	1.31	− 1.37	− 0.11	0.26	2.01	Extremely difficult
Item24	− 3.36	0.15	− 1.71	0.45	0.21	1.89	Difficult
Item25	− 0.88	− 0.05	− 0.08	3.88	0.01	0.23	Moderately difficult
Item26	− 2.63	− 2.13	− 0.39	− 0.33	0.30	1.20	Difficult
Item27	− 3.73	− 0.96	1.98	− 0.06	0.13	1.69	Difficult
Item28	− 1.48	− 2.40	0.88	1.39	0.05	0.51	Difficult
Item29	− 1.37	− 2.21	0.35	− 1.80	0.11	0.48	Moderately difficult
Item30	− 0.45	− 0.36	− 0.93	0.18	0.19	0.44	Moderately difficult
Item31	− 1.31	0.30	0.55	0.37	0.17	1.80	Difficult
Item32	0.05	− 0.59	− 0.53	− 0.26	0.00	− 0.06	Easy
Item33	0.53	− 0.67	− 1.36	− 0.63	0.00	− 0.32	Easy
Item34	0.39	− 0.56	− 1.17	− 1.34	0.07	− 0.21	Easy
Item35	− 0.59	− 1.39	− 2.15	0.45	0.21	0.23	Moderately difficult

In addition to this, Table 4 column 5 contains a parameter c (lower asymptote), which represents the probability that the item will be correct when guessing alone based solely on the item. In fact, by definition, the value of c is independent of the level of ability, so its value will not differ based on the level of ability. Consequently, regardless of the level of ability of the examinee, they will have exactly the same

Table 4 Three-multidimensional item discriminating parameter of physics tests

Item	a1	a2	a3	g	Remarks	MDISC	Remarks
Item1	0.64	1.01	2.64	0.12	Acceptable	2.90	VHD
Item2	− 0.38	− 1.73	0.87	0.01	Acceptable	1.97	VHD
Item3	− 2.78	− 0.64	− 1.14	0.18	Acceptable	3.07	VHD
Item4	− 2.96	− 0.95	− 0.34	0.20	Acceptable	3.13	VHD
Item5	− 4.01	− 0.59	− 0.31	0.26	Acceptable	4.06	VHD
Item6	− 1.26	− 1.35	− 0.60	0.26	Acceptable	1.94	VHD
Item7	0.25	0.91	1.66	0.14	Acceptable	1.91	VHD
Item8	− 0.09	− 1.77	1.18	0.12	Acceptable	2.13	VHD
Item9	− 1.06	− 2.52	1.40	0.12	Acceptable	3.07	VHD
Item10	− 1.34	− 0.66	2.21	0.00	Acceptable	2.67	VHD
Item11	− 1.50	0.04	1.81	0.02	Acceptable	2.35	VHD
Item12	− 0.48	− 1.97	0.48	0.15	Acceptable	2.08	VHD
Item13	− 0.76	− 1.43	− 0.21	0.20	Acceptable	1.63	HD
Item14	− 0.75	− 1.73	− 0.12	0.40	Not acceptable	1.89	VHD
Item15	− 0.84	− 1.75	− 2.05	0.22	Acceptable	2.83	VHD
Item16	0.15	0.44	0.02	0.35	Acceptable	0.47	LD
Item17	− 1.55	− 1.30	0.19	0.34	Acceptable	2.03	VHD
Item18	0.63	2.01	1.01	0.33	Acceptable	2.33	VHD
Item19	− 2.21	− 1.33	0.63	0.38	Not acceptable	2.66	VHD
Item20	0.64	− 2.72	− 0.66	0.29	Acceptable	2.87	VHD
Item21	1.02	− 1.16	− 1.81	0.37	Not acceptable	2.38	VHD
Item22	1.00	− 2.14	− 0.48	0.22	Acceptable	2.41	VHD
Item23	1.31	− 1.37	− 0.11	0.26	Acceptable	1.90	VHD
Item24	0.15	− 1.71	0.45	0.21	Acceptable	1.78	VHD
Item25	− 0.05	− 0.08	3.88	0.01	Acceptable	3.88	VHD
Item26	− 2.13	− 0.39	− 0.33	0.30	Acceptable	2.19	VHD
Item27	− 0.96	1.98	− 0.06	0.13	Acceptable	2.20	VHD
Item28	− 2.40	0.88	1.39	0.05	Acceptable	2.91	VHD
Item29	− 2.21	0.35	− 1.80	0.11	Acceptable	2.87	VHD
Item30	− 0.36	− 0.93	0.18	0.19	Acceptable	1.01	MD
Item31	0.30	0.55	0.37	0.17	Acceptable	0.73	MD
Item32	− 0.59	− 0.53	− 0.26	0.00	Acceptable	0.83	MD
Item33	− 0.67	− 1.36	− 0.63	0.00	Acceptable	1.64	HD
Item34	− 0.56	− 1.17	− 1.34	0.07	Acceptable	1.87	VHD
Item35	− 1.39	− 2.15	0.45	0.21	Acceptable	2.60	VHD

chance of getting the item correct by guessing the response (Baker, 2017). As the parameter c has a theoretical range of $0 \leq c \leq 1.0$, in practice it is considered unacceptable to have values greater than 0.35 regardless of how many times it is tested. Consequently, Baker (2001) found the range between $0 \leq c \leq 0.35$ to be acceptable and recommended that it be used. In this regard, the guessing indices as depicted

in Table 3 demonstrated that the majority of items have a guessing index within the acceptable range of 0.35 or less, except for items (14, 19 and 21) that have a lower-asymptote index that exceeds 0.35. This could indicate that there are not many items in the physics test that may be susceptible to guessing behaviour.

Discussion

The Item Response Theory (IRT) framework emphasizes the importance of considering test dimensionality during test development and establishing a strong validity foundation to guide the construction of effective tests. Test dimensionality refers to the number of traits or dimensions the test aims to measure. Therefore, test items are designed and organized to align with these intended traits or dimensions. In statistical dimensionality analysis, the goal is to identify the underlying constructs of the test instrument, providing evidence of test validity. Many testing programs worldwide typically assume a simple structure when analyzing data, meaning that each item measures only one trait. However, some testing programs have recognized the need for more complex analyses beyond simple structures, particularly in the context of Multidimensional Item Response Theory (MIRT) modeling. This growing field requires dimensionality analyses tailored to MIRT models. This study focuses specifically on the measurement instruments used at NOUN, particularly the PHY 101 physics test, which contains items described as multidimensional. These items require examinees to demonstrate knowledge across multiple content types and cognitive ability levels. Consequently, accurately estimating item parameters becomes crucial in assessing an examinee's ability in this subject, directly affecting the quality of test items. To address this, a multidimensional approach was employed to estimate the item parameters of the PHY 101 physics test items using the mirt package in the R language for statistical computing. The dimensionality analysis results led to the selection of the M3PL IRT model for calibrating PHY 101 test items based on model fit. These findings support previous research by Ackerman et al. (2003), Tobih et al. (2023), Liu et al. (2013), and Ul Hassan and Miller (2022), suggesting that increased dimensionality and complexity result in better model fit. Complex models with more dimensions are considered superior when grouping similar questions or items. This aligns with Wiberg's (2012) observation that item parameters and ability estimates tend to align with the strongest factor in a multidimensional dataset with strong factors beyond the primary unidimensional parameterization. Therefore, MIRT is more suitable for real-world applications where multidimensionality is realistic.

Additionally, the difficulty indices of the physics test items were found to be of high quality. This suggests that the test is most suitable for assessing the proficiency of individuals with moderate to high proficiency in physics, but not for those with low proficiency. Some items were challenging, while others were moderately difficult. These findings are consistent with research by Adekunle et al. (2021), which highlight the impact of test length on item statistics and the precision of item difficulty parameters concerning test dimensions. In contrast, Zulaeha et al. (2020) suggested that a test length of 15 displayed a median correlation of 0.78 for variance estimates, indicating that the questions were more difficult, potentially leading to

guessing by respondents. Additionally, item difficulty parameters were highly influenced by sample size.

Regarding the quality of discrimination item parameters, the study found that the physics test items effectively discriminated between proficient and non-proficient examinees. This aligns with the findings of Kose and Demirtasli (2012), Zulaeha et al. (2020), Liu et al. (2021), Zhang (2012), and Adekunle et al. (2021), which suggest that discrimination parameters are associated with multiple dimensions. Furthermore, as argued by Liu et al. (2013) and Mark et al. (1983), an item's discrimination power on one dimension depends on its ability on other dimensions. Therefore, a dimension's discriminating power increases as the abilities on other dimensions increase.

Additionally, the results indicated that test-takers were generally not influenced by chance factors when responding to test items, except for a few items (e.g., items 14, 19, and 21 with c-parameters of 0.40, 0.38, and 0.37). The difficulty of these items may lead to guessing behavior rather than reflecting true ability, and other factors beyond proficiency may affect students' responses. Therefore, employing an appropriate multidimensional model to determine the c parameter, rather than a unidimensional model, is crucial to avoid overestimating test item indices. Past research has shown that some methods of estimating MIRT item parameters can be influenced by the presence of chance factors.

Conclusion

Researchers who are interested in estimating MIRT model item parameters may find the results of this study useful. This research sought to examine various factors associated with NOUN's physics test, rather than focusing on the unidimensional structure, where individual items were associated with only one latent trait. Thus, these results might provide practitioners with additional insights into MIRT data modeling. The results presented above suggest that parameter estimates obtained using the fitted M3PL model to the NOUN physics test remarked quality item parameters (difficulty, discrimination, and chance factor) with reduced bias. The study concludes that multidimensional models of IRT are dependable and most appropriate for evaluating the psychometric quality of NOUN physics test items since the test evaluated test-takers in various areas (e.g., cognitive processing/reasoning, reading skills and problem-solving skills). According to the study, the university's directorate of examination and assessment needs to establish a psychometric unit and hire assessment experts with a deep understanding of test theories to help validate the instrument before administration. Establish item parameters based on multidimensional models for courses with inherent variables that affect students' performance and train and retrain lecturers who develop these items to keep them current. IRT practitioners working with multidimensional data should be aware of the implications presented above when responding to IRT questions that do not measure a single trait. As a result, parameter estimation for items with multidimensional structures will be generally less accurate than for items with unidimensional structures and neglecting multidimensional structure will almost certainly result in biased difficulty, discrimination, and guesswork. Research in the future should address

limitations of this study. For calibration using R programming software, the time to converge was longer. Moreover, this study did not consider establishing examinees' ability, which may be a potential area of concern and focus for future research. It is also possible to establish other NOUN courses parameters to assess whether the model used to model test-takers' responses is appropriate, which can affect their performance negatively or positively.

Acknowledgements 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.

Author contributions Musa Adekunle, Ayanwale handled the following: Conceptualisation, writing—original draft preparation, methodology, data analysis, visualization, discussion and conclusion. Jamiu Oluwadamilare, Amusa handled the following: resources, writing—review and editing and supervision. Adekunle Ibrahim, Oladejo handled the following: Data sorting, data curation and references alignment. Funmilayo, Ayedun handled the following: Formatting of the manuscript.

Funding Open access funding provided by University of Johannesburg. This research was supported by senate research grants (Ref: NOUN/DRA/LARTL/005/VOL1) from the National Open University of Nigeria (NOUN), Abuja. However, the article should be published as closed access.

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

Declarations

Conflicts of interest The authors declare no conflict of interest.

References

Ackerman, T. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement*, *20*(4), 311–329. https://doi.org/10.1177/014662169602000402.

Ackerman, T. A. (2010). The theory and practice of Item Response Theory by De Ayala. *R J Journal of Educational Measurement*, *47*(4), 471–476. https://doi.org/10.1111/j.1745-3984.2010.00124.x.

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using Multidimensional Item Response Theory to Evaluate Educational and Psychological tests. *Educational Measurement: Issues and Practice*, *22*(3), 37–51. https://doi.org/10.1111/j.1745-3992.2003.tb00136.x.

Adekunle, F. T., Oluwafemi, A. O., & Afolabi, E. R. I. (2021). Psychometric Properties of Geography in Osun State Unified Promotion Examinations using Multidimensional Item Response Theory. *Nigerian Journal of Educational Research and Evaluation*, *20*, 29–47. http://www.journal.theasseren.org.ng/index.php/naere/issue/archive.

Ajeigbe, T. O., & Afolabi, E. R. I. (2014). Assessing unidimensionality and differential item functioning in qualifying examination for senior secondary school students, Osun State, Nigeria. World Journal of Education, 4 (4). https://www.sciedu.ca/journal/index.php/wje/article/view/5086.

Akinboboye, J. T., & Ayanwale, M. A. (2021). Bloom taxonomy usage and psychometric analysis of classroom teacher made test. *African Multidisciplinary Journal of Development*, *10*(1), 10–21.

Algina, J., & Swaminathan, H. (2015). Psychometrics: Classical test theory. *International Encyclopedia of the Social & Behavioral Sciences: Second Edition* (pp. 423–430). Elsevier Inc. https://doi.org/10.1016/B978-0-08-097086-8.42070-2.

Amusa, J. O., Ayanwale, M. A., Oladejo, I. A., & Ayedun, F. (2022). Undergraduate physics test dimensionality and conditional independence: Perspective from latent traits model (ltm) Package of R Language. *International Journal of Assessment and Evaluation*, *29*(2), 47–61. https://doi.org/10.18848/2327-7920/CGP/v29i02/47-61.

Ayanwale, M. A. (2019). Efficacy of Item Response Theory in the Validation and Score Ranking of Dichotomous and Polytomous Response Mathematics Achievement Tests in Osun State, Nigeria. In *Doctoral Thesis, Institute of Education, University of Ibadan* (Issue April). https://doi.org/10.13140/RG.2.2.17461.22247.

Ayanwale, M. A. (2021). Calibration of Polytomous Response mathematics Achievement Test using generalized partial credit model of Item Response Theory. *EDUCATUM Journal of Science Mathematics and Technology*, *8*(1), 57–69. https://doi.org/10.37134/ejsmt.vol8.1.7.2021.

Ayanwale, M. A. (2023). Test score equating of multiple-choice mathematics items: Techniques from characteristic curve of modern psychometric theory. *Discov Educ*, *2*, 30. https://doi.org/10.1007/s44217-023-00052-z.

Ayanwale, M. A., & Adeleke, J. O. (2020). Efficacy of Item Response Theory in the Validation and Score Ranking of Dichotomous Response Mathematics Achievement Test. Bulgarian Journal of Science and Education Policy, 14 (2), 260–285. http://bjsep.org/.

Ayanwale, M. A., & Ndlovu, M. (2021). Ensuring scalability of a cognitive multiple-choice test through the Mokken Package in R Programming Language. *Education Sciences*, *11*(12), 794. https://doi.org/10.3390/educsci11120794.

Ayanwale, M. A., Adeleke, J. O., & Mamadelo, T. I. (2018). An Assessment of Item Statistics Estimates of Basic Education Certificate Examination through Classical Test Theory and Item Response Theory approach. *International Journal of Educational Research Review*, *3*(4), 55–67. https://doi.org/10.24331/ijere.452555.

Ayanwale, M. A., Adeleke, J. O., & Mamadelo, T. I. (2019). Invariance person estimate of Basic Education Certificate examination: Classical test theory and item response theory scoring perspective. *Journal of the International Society for Teacher Education*, *23*(1), 18–26. https://isfte.org/jisteour-bi-annual-journal/journal%20volumes/publicly-accessibly-abstracts-only/.

Ayanwale, M. A., Isaac-Oloniyo, F. O., & Abayomi, F. R. (2020). Dimensionality Assessment of Binary response test items: A non-parametric Approach of bayesian item response theory measurement. *International Journal of Evaluation and Research in Education*, *9*(2), 412–420. https://doi.org/10.11591/ijere.v9i2.20454.

Ayanwale, M. A., Chere-Masopha, J., & Morena, M. (2022). The classical test or item response measurement theory: The Status of the Framework at the Examination Council of Lesotho. *International Journal of Learning Teaching and Educational Research*, *21*(8), 384–406. https://doi.org/10.26803/ijlter.21.8.22.

Baker, F. B. (2001). *The basics of Item Response Theory. Test Calibration*. ERIC Clearinghouse on Assessment and Evaluation.

Baker, F. B., & Kim, S. (2017). *The Basics of Item Response Theory Using R* (S. E. Fienberg (Ed.)). Springer International Publishing. https://doi.org/10.1007/978-3-319-54205-8_1.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and non-compensatory multidimensional item response models using Markov Chain Monte Carlo. *Applied Psychological Measurement*, *27*(6), 395–514. https://doi.org/10.1177/0146621603258350.

Chalmers, R., P (2012). Mirt: A Multidimensional Item Response Theory Package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06.

De Champlain, A. F., & Gessaroli, M. E. (1998). Assessing the dimensionality of item response matrices with small sample sizes and short test lengths. *Applied Measurement in Education*, *11*, 231–253. https://Doi/10.1.1.899.504.

De La Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, *30*(3), 295–311. https://doi.org/10.3102/10769986030003295.

Drasgow, F., & Parsons, C. K. (1983). Application of Unidimensional Item Response Theory models to Multidimensional Data. *Applied Psychological Measurement*, *7*(2), 189–199. https://doi.org/10.1177/014662168300700207.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Finch, H. (2011). Multidimensional item response theory parameter estimation with nonsimple structure items. *Applied Psychological Measurement*, *35*(1), 67–82. https://doi.org/10.1177/0146621610367787.

Finkelman, M., Hooker, G., & Wang, Z. (2010). Prevalence and magnitude of paradoxical results in multidimensional item response theory. *Journal of Educational and Behavioral Statistics*, *35*(6), 744–761. https://doi.org/10.3102/1076998610381402.

Gessaroli, M. E., & De Champlain, A. F. (1996). Using an approximate chi-square statistic to test the number of dimensions underlying the responses to a set of items. *Journal of Educational Measurement*, *33*, 157–179. https://www.jstor.org/stable/1435181.

Ha, D. T. (2017). Applying Multidimensional three-parameter logistic model (M3PL) in validating a multiple-choice test. *International Journal of Scientific and Research Publications*, *7*(2), 175–183.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Kluwer-Nijhoff.

Hasmy, A. (2014). Compare unidimensional& multidimensional Rasch model for test with multidimensional construct and items local dependence. *Journal of Education and Learning*, *8*(3), 187–194.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*(2), 139–164. https://doi.org/10.1177/014662168500900204.

Hattie, J., Krakowski, K., Rogers, H. J., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement*, *20*, 1–14. https://doi.org/10.1177/014662169602000101.

Hooker, G. (2010). On separable test, correlated priors, and paradoxical results in multidimensional item response theory. *Psychometrika*, *75*(4), 694–707. https://doi.org/10.1007/s11336-010-9181-5.

Hooker, G., Finkelman, M., & Schwartzman, A. (2009). Paradoxical results in multidimensional item response theory. *Psychometrika*, *74*(3), 419–442. https://doi.org/10.1007/s11336-009-9111-6.

Immekus, J. C., Snyder, K. E., & Ralston, P. A. (2019). Multidimensional item response theory for factor structure assessment in educational psychology research. *Frontiers in Education*. https://doi.org/10.3389/feduc.2019.00045

Jang, E. E., & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *Journal of Educational Measurement*, *44*(1), 1–21. https://doi.org/10.1111/j.1745-3984.2007.00024.x.

Jordan, P., & Spiess, M. (2012). Generalization of paradoxical results in multidimensional item response theory. *Psychometrika*, *77*(1), 127–152. https://doi.org/10.1007/s11336-011-9243-3.

Jordan, P., & Spiess, M. (2018). A new explanation and proof of the paradoxical scoring results in multidimensional item response models. *Psychometrika*, *83*(4), 831–846. https://doi.org/10.1007/s11336-017-9588-3.

Kose, I. A., & Demirtasli, N. C. (2012). Comparison of Unidimensional and Multidimensional models based on Item Response Theory in terms of both variables of test length and sample size. *Procedia - Social and Behavioral Sciences*, *46*, 135–140. https://doi.org/10.1016/j.sbspro.2012.05.082.

Li, Y., Jiao, H., & Lissitz, R. (2012). Applying Multidimensional Item Response Theory models in validating test dimensionality: An example of K-12 large-Scale Science Assessment. *Journal of Applied Testing Technology*, *13*(2), 220–239.

Liu, H. Y., Luo, F., Wang, Y., & Zhang, Y. (2013). Item parameter estimation for Multidimensional Measurement: Comparisons of SEM and MIRT Based methods. *Acta Psychologica Sinica*, *44*(1), 121–132. https://doi.org/10.3724/sp.j.1041.2012.00121.

Lord, F. M. (1980). Application of item response theory to practice testing problems. Lawrence Erlbaum Associates.

Mark, D., Robert, L., & McKinley, M. D. (1983). The definition of Difficulty and discrimination for Multidimensional Item Response Theory models. Education Resources Information Center, 2–14. https://www.researchgate.net/publication/234738229.

Mulaik, S. A. (1972). A mathematical investigation of some multidimensional Rasch model for psychological tests. Paper presented at the annual meeting of the Psychometric Society, Princeton, New York.

Ojerinde, D., Popoola, B., Ojo, F., & Onyeneho, P. (2012). *Introduction to Item Response Theory: Parameter models, estimation and application*. Marvelouse Mike Press Ltd.

Owolabi, T., Akintoye, H., Amusa, J. O., & Ayanwale, M. A. (2023). Educational testing techniques in senior secondary school physics in Nigeria: are we ascertaining the development of requisite behavioural objectives? African Perspectives of Research in Teaching & Learning (APORTAL), (1).

Ozdemir, B., & Gelbal, S. (2022). Measuring language ability of students with compensatory multidimensional CAT: A post-hoc simulation study. *Education and Information Technologies*. https://doi.org/10.1007/s10639-021-10853-0.

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*, 401–412. https://doi.org/10.1177/014662168500900409.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). New York: Springer-Verlag.

Reckase, M. D. (2007). Multidimensional item response theory. In C. R. Rao, & S. Sinharay (Eds.), Handbook of statistics and psychometrics (pp. 607–642). Elsevier.

Reckase, M. D. (2009). *Multidimensional Item Response Theory (Statistics for Social and Behavioral Sciences)*. http://www.amazon.com/Multidimensional-Response-Statistics-Behavioral-Sciences/dp/0387899758/ref=sr_1_1?ie=UTF8&qid=1363871688&sr=8-1&keywords=Multidimensional+Item+Response+Theory+(Statistics+for+Social+and+Behavioral+Sciences).

Robitzsch, A. (2015). *Package sirt. Supplementary Item Response Theory Models*. https://www.google.com/search?q=Supplementary+Item+Response+Theory+Models+(sirt)+package+(Robitzsch%2 C+2019)&oq=Supplementary+Item+Response+Theory+Models+(sirt)+package+(Robitzsch%2 C+2019)&aqs=chrome.69i57.38699j0j7&sourceid=chrome&ie=UTF-8.

Robitzsch, A. (2020). *Supplementary Item Response Theory Models (sirt)*. https://search.r-project.org/CRAN/refmans/sirt/html/sirt-package.html.

Seock-Ho, K. B. (2004). Item Response Theory: Parameter Estimation Techniques. In *Biometrics* (Vol. 50, Issue 3). Marcel Dekker. https://doi.org/10.2307/2532822.

Sheng, Y., & Wikle, C. K. (2007). Comparing multiunidimensional and unidimensional item response theory models. *Educational and Psychological Measurement*, *67*(6), 899–919. https://doi.org/10.1177/0013164406296977.

Sijtsma, K., & Junker, B. W. (2006). Item response theory: Past performance, Present Developments, and future expectations. *Behaviormetrika*, *33*(1), 75–102. https://doi.org/10.2333/bhmk.33.75.

Smith, J. (2009). Some issues in item response theory: Dimensionality assessment and models for guessing. Unpublished Doctoral Dissertation. University of South California.

Spencer, S. G. (2004). The strength of multidimensional item response theory in exploring construct space that is multidimensional and correlated. Ph. D. thesis., Doctoral Dissertation, Brigham Young University-Provo. https://scholarsarchive.byu.edu/etd/224/.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrik*, *52*, 589–611.

Stout, W., Habing, B., Douglas, J., & Kim, H. R. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, *20*, 331–354.

Svetina, D., & Levy, R. (2014). A Framework for Dimensionality Assessment for Multidimensional Item Response models. *Educational Assessment*, *19*(1), 35–57. https://doi.org/10.1080/10627197.2014.869450.

Sympson, J. B. (1978). A model for testing with multidimensional items. In Weiss D.J. (Ed.). Proceeding of the 1977 Computerized Adaptive Testing Conference, University of Minnesota, Minneapolis.

Team, R. (2019). *A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. http://www.r-project.org.

Tobih, D. O., Ayanwale, M. A., Ajayi, O. A., & Bolaji, M. V. (2023). The use of measurement frameworks to explore the qualities of test items. *Int J Eval & Res Educ*, *12*(2). https://doi.org/10.11591/ijere.v12i2.23747.

Ul Hassan, M., & Miller, F. (2022). Discrimination with unidimensional and multidimensional item response theory models for educational data. *Communications in Statistics: Simulation and Computation*, *51*(6), 2992–3012. https://doi.org/10.1080/03610918.2019.1705344.

Walker, C. M., & Beretvas, S. N. (2000). Using multidimensional versus unidimensional ability estimates to determine student proficiency in mathematics. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Wiberg, M. (2012). Can a multidimensional test be evaluated with unidimensional item response theory? *Educational Research and Evaluation*, *18*(4), 307–320. https://doi.org/10.1080/13803611.2012.670416.

Yang, S. (2007). A comparison of unidimensional and multidimensional Rasch models using parameter estimates and fit indices when assumption of unidimensionality is violated. Ph. D. thesis, doctoral dissertation, The Ohio State University.

Zhang, J. (2004). Comparison of unidimensional and multidimensional approaches to irt parameter estimation. *ETS Research Report Series*, *24*(2), 1–40. https://doi.org/10.1002/j.2333-8504.2004.tb01971.x.

Zhang, J. (2007). Conditional covariance theory and DETECT for polytomous items. *Psychometrika*, *72*(1), 69–91. https://doi.org/10.1007/s11336-004-1257-7.

Zhang, J. (2013). A procedure for dimensionality analyses of response data from various test designs. *Psychometrika*, *78*(1), 37–58. https://doi.org/10.1007/s11336-012-9287-z.

Zhang, B., & Stone, C. A. (2008). Evaluating item fit for multidimensional item response models. *Educational and Psychological Measurement*, *68*, 181–196. https://doi.org/10.1177/0013164407301547.

Zhang, J., & Stout, W. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika*, *64*, 129–152.

Zhang, J., & Stout, W. (1999b). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*, 213–249.

Zulaeha, O., Rahayu, W., & Sastrawijaya, Y. (2020). The estimates item parameter for Multidimensional Three-Parameter Logistics. *KnE Social Sciences*, *2020*, 315–322. https://doi.org/10.18502/kss.v4i14.7889.

Authors and Affiliations

Musa Adekunle Ayanwale[1] · Jamiu Oluwadamilare Amusa[2] · Adekunle Ibrahim Oladejo[3] · Funmilayo Ayedun[2]

✉ Musa Adekunle Ayanwale
ayanwalea@uj.ac.za

Jamiu Oluwadamilare Amusa
jamusa@noun.edu.ng

Adekunle Ibrahim Oladejo
gbadegeshin86@gmail.com

Funmilayo Ayedun
fayedun@noun.edu.ng

[1] Department of Science and Technology Education, University of Johannesburg, Johannesburg, South Africa

[2] Department of Science Education, National Open University of Nigeria, Abuja, Nigeria

[3] Africa Centre of Excellence for Innovative and Transformative STEM Education, Lagos, Nigeria