Research in Social Sciences and Technology

https://ressat.org

E-ISSN: 2468-6891

Volume: 9 Issue: 3 2024

pp. 387-401

Assessment of Item Parameter and Model-Data Fit of 4IR Chemistry Teachers' Effectiveness Scale in Secondary Schools

Chidubem Deborah Adamu[a]

***** Corresponding author
Email: cadamu@uj.ac.za

a. Department of Education Leadership and Management, Faculty of Education, University of Johannesburg, South Africa

10.46303/ressat.2024.66

Article Info
Received: June 11, 2024
Accepted: September 13, 2024
Published: October 21, 2024

How to cite
Adamu, C. D. (2024). Assessment of Item Parameter and Model-Data Fit of 4IR Chemistry Teachers' Effectiveness Scale in Secondary Schools. *Research in Social Sciences and Technology*, *9*(3), 387-401. https://doi.org/10.46303/ressat.2024.66

ABSTRACT

The Fourth Industrial Revolution Chemistry Teachers Effectiveness Scale (4IRCTES) was evaluated in secondary schools in Southwest Nigeria to determine its item discrimination, item parameters, and model-data fit. This study utilised a descriptive survey research design and included 4,986 Chemistry teachers in the southwestern region of Nigeria, with a sample of 35 Chemistry teachers. The 4IRCTES instrument was used to collect data. Research question one focused on item discrimination using the Multidimensional Graded Response Model (MGRM) of the Item Response Theory (IRT). Research question two examined the overall model-data fit, or M2 statistic. Results indicated that the 4IRCTES items effectively discriminated between teachers with low and high effectiveness. Also, result showed that the MGRM of the IRT is the substantial model-data fit for the 4IRCTES. The study concluded that the MGRM of the IRT provided a strong model-data fit for the 4IRCTES, and that the scale effectively distinguished between ineffective and effective teachers.

KEYWORDS

Chemistry Teachers Effectiveness Scale (CTES); factor; Fourth Industrial Revolution (4IR); highly effective teachers; item discrimination; model-data fit.

INTRODUCTION

Background to the Study

The Fourth Industrial Revolution (4IR) is currently transforming the way education is delivered. In both developed and developing countries, traditional home instruction and coaching are being replaced by this revolution. Nowadays, students are learning online through various computer programs such as Zoom, WhatsApp, and YouTube. While classroom instruction remains important, educators are incorporating interactive online learning platforms like Google Classroom, Kahoot, Zoom Education, Seesaw, Photomath, Edmodo, Prezi, Thinglink, Class Dojo, Quizlet, Storyboard, Animoto, Educreations, and others to enhance their students' learning outcomes. Although the 4IR has not completely altered the nature of schooling yet, it is significantly impacting the way education is delivered.

In order for the 4IR to thrive in Nigeria's education system, educators at all levels need to be proficient in technology. The reliance on traditional teaching methods must be minimal. Teachers must understand that lectures alone are not always engaging for students. Students are more motivated when presented with a variety of modern teaching methods to enhance their learning experience. By incorporating contemporary teaching methods, technology can help foster new teacher-student relationships and facilitate the transfer of knowledge from teachers to students. To embrace modern pedagogy, many secondary school administrators in Nigeria are now encouraging their Grade 10 students to bring smartphones and laptops to class for computer science lessons.

The field of education has undergone significant changes in recent years, with a greater focus on accountability and a thorough examination of the factors that influence educational outcomes. Various elements, including teacher effectiveness, play a crucial role in determining how well high school students perform on both school-based and standardized achievement tests. Additionally, students' learning is impacted by a multitude of factors such as peer interactions, family support, home environment, school resources, community involvement, leadership, and the overall school environment (Little, 2009).

Research has consistently shown that teachers have the most significant impact on students' academic success over time (Stronge, 2018). To better understand how teachers influence student learning, researchers have begun to analyse the specific characteristics and instructional strategies utilised by top educators. Recognising the pivotal roles that exceptional teachers play in shaping students' learning experiences is fundamental to evaluating the quality of education.

The importance of teachers as a fundamental component cannot be overstated in ensuring quality in the teaching and learning of a subject at all levels of education, including Chemistry education in secondary schools. According to Darling-Hammond (2015), schools are only as strong as their educators, who are vital to the education system. Teachers, as creators of knowledge, play a crucial role in the educational process. Teaching in the classroom is just one aspect of a teacher's role in many countries. A good teacher goes beyond identifying issues in students' homes, communities, and psychological needs and providing solutions. Therefore, the ability of teachers to exceed expectations and improve students' success and test scores is a measure of their effectiveness.

There has been a decline in the performance of secondary school students in Chemistry, both in the classroom and on standardized tests. As a result, parents and other education stakeholders are starting to question the effectiveness of Chemistry teachers in the 4IR. To address this issue, the researcher developed the 4IRCTES. The aim of the study was to assess the model-data fit and item parameters, particularly item discrimination, of the 4IRCTES in secondary schools in Southwest Nigeria.

Physics, Chemistry, and Biology are the three most popular science subjects in Nigeria, with Chemistry playing a crucial role in secondary science education. This is why most students majoring in Physics or Biology also choose Chemistry as a minor. Chemistry is considered a fundamental subject in science education, with its contributions to the advancement of science and technology well-documented (Hassan & Salihu, 2019). In the Nigerian educational system, Chemistry is a prerequisite for natural science courses and other science subjects (Adesoji, 2008; Edomwonyi & Avaa, 2011; Bugaje, 2013). Chemistry is a hands-on subject that requires highly effective teachers for proper instruction. This ensures that teachers can produce students who are skilled and relevant in practical Chemistry. However, this study focuses not on Chemistry as a subject, but on the effectiveness of the teachers teaching Chemistry.

In order to be highly effective in practical Chemistry, a teacher of the subject must be capable of teaching both high-ability and at-risk students (Cline & Schwartz, 1999; Kaul, 2015; Siegle et al., 2016). This is crucial because even students with high capabilities can face academic challenges. Similarly, students who are at risk of failing in school can also be exceptionally gifted. However, students who require remediation are at risk of retention or dropping out. They show low self-efficacy, or have inadequate academic skills, and are generally referred to as "at-risk" students. Teachers face numerous behavioural and instructional challenges with these students, but it is essential to provide them with strong support to help them become effective learners and overcome their current circumstances (Sagor & Cox, 2013). These students often experience educational disparities due to differences in skin tone, ethnicity, language, and social status.

Students with above-average abilities are known as highly-able students. They are recognised for their exceptional intelligence, creativity, independence, critical thinking skills, leadership qualities, sensitivity, and curiosity. Teachers of these students must consider their unique needs and characteristics, just as they do for all students (Van Tassel-Baska & Hubbard, 2016). This study focuses on the item parameter of 4IRCTES, which examines whether highly capable Chemistry teachers can differentiate themselves from less skilled peers in terms of the effectiveness of their instruction when using technology. However, special attention is given to item discrimination.

Item analysis is the process of evaluating test items in educational testing to assess the overall quality of the test as well as the quality of individual items. Examinee responses to specific items are analysed as part of the item analysis process, which assesses both the overall quality of the test and the quality of individual items. Item Response Theory (IRT) is used to estimate item parameters throughout the item analysis process. More reliable statistical techniques and analytical tools compatible with 4IR are employed for item analysis. Teachers in postsecondary education institutions are also researchers. It is expected that these educators

have varying levels of expertise in statistical analysis related to their research. Often, researchers lack a basic understanding of statistical techniques and analytical tools, leading them to outsource data analysis for their studies. Therefore, the main objective of this study was to evaluate the fit of the 4IRCTES items to their underlying construct, as well as the reliability of the statistical techniques and analytical tools used to investigate the item discrimination parameter.

THEORETICAL REVIEW

The Holy Bible recognises the importance of a trustworthy scale for measurement in research. The Bible states that the Lord loves precise weights but hates the use of dishonest scales (Proverbs 11:1, New Living Translation). A trustworthy scale works hand in hand with an exact weight, making it essential that a scale's items measure what they are supposed to measure. With the use of modern technology, researchers and psychometricians can evaluate the model-data fit of a scale by factoring its items using Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). While CFA confirms the factor structure of a collection of observable variables, EFA assists in determining the number of factors among the items in a scale and which factors are determined by which items.

Although, the majority of early users of IRT models were in education, IRT models are still commonly employed in the social sciences today, particularly for developing new psychological notions. It is evident that item response theory is helpful for developing scales and estimating latent traits in a variety of research areas (Loken & Rukison, 2010). Item discrimination (a), item difficulty (b), and the guessing parameter comprise the majority of these statistical indicators (Lotobi & Basil, 2019). There are One-, Two-, and Three-Parameter Logistic Models (1-PL, 2-PL, and 3-PL, respectively) that are parametric variations of IRT models. The examinee's likelihood of providing the right answer to a question is expressed by the complete 3-item parameter logistic (3-PL) model (Galdin & Laurencelle, 2010). Equation (1) presents it as follows:

$$Pj\ (\theta_r) = c_j\ + (1 - c_j\ \frac{1}{1+e^{-a_j(\theta_r - b_j)}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

The parameter θ represents the test-takers' individual ability level in the equation above. The parameter bj is the difficulty parameter, explaining how challenging it is for test-takers to answer each item correctly. Parameter aj is the discrimination parameter, explaining how well test items can differentiate between test-takers with high ability and those with low ability. Parameter Cj is the guessing parameter, indicating the likelihood that a test-taker with low ability will correctly guess an item. The item difficulty parameter (bj) ranges from -1.0 to +1.0, discrimination parameters (aj) are above 0.30, and guessing parameters (cj) are below 0.25 (for items with four options) and 0.20 (for items with five options). These values are based on calibration methods described by Omorogiuwa (2009) and Guyer and Thompson (2011), as cited in Ethe and Odjegba (2019).

Items that meet the IRT statistical requirements are automatically selected by X-Calibre, which also calibrates the item parameters. The 1-PL, 2-PL, and 3-PL versions of the IRT correspond to the parameters bj, aj, and cj respectively. Additionally, there is a Four-Parameter Logistic (4-PL) IRT model. In this study, the latent component (θ, or effectiveness) in 4IRCTES

was measured using the 2-PL of the IRT model, which includes the instructional ability and item discrimination of Chemistry teachers (Samejima, 1969).

The Graded Response Model (GRM) is one of the IRT models used to measure polytomously scored multidimensional items. Other models include the Nominal Response Model (NRM), Partial Credit Model (PCM), Generalized Partial Credit Model (GPCM), and Rating Scale Model (RSM). The effectiveness scale for fourth industrial revolution chemistry teachers utilised a polytomous scoring system due to the latent construct of teaching effectiveness having multiple possible outcomes. Multidimensionality arises in a scale when items assess two or more latent components, challenging the unidimensionality assumption of IRT.

IRT serves as a framework for assessing how well each test item functions in an assessment (Ethe & Odjegba, 2019). Equation (1) shows the item discrimination, or how well an item distinguishes between people with different levels of θ or ability, as represented by the a-parameter or slope estimate. Items with low slope—those that are almost at zero—are problematic because they fail to distinguish between teachers who are effective at different levels. Additionally, given the respondents' level of θ (effectiveness), each b parameter or severity estimate (b1, b2, b3, b4) in equation (1) indicates the point along θ where one response category becomes more likely to be endorsed than any other alternative. Items with evenly dispersed b values over the θ range allow for clear differentiation of individuals with different levels of θ. Items with excessive b values (more than 4.5 standard deviations in either direction) or that are too close together are not as desirable because it is unclear what an individual's degree of θ is based on the selected response option.

In the realm of educational testing and measurement, two theories are considered helpful in scale development: IRT and Classical Test Theory (CTT). For this investigation, IRT was used as the theoretical foundation. The theory gained traction because it emphasised the connection between each item and the fundamental idea or skill that the instrument was designed to evaluate. Information regarding the process of evaluating an item's properties (discrimination, difficulty, and guessing) can be found in an IRT-developed test. Classical test theory, a conventional approach to test theory, is limited to a true score theory and, unlike IRT, applies statistical tools and mathematical models from physical measurement to problems that were thought to be equivalent in mental measurement (Embretson & Reise, 2010). CTT has been criticised for its real score, which is said to depend on the test's content rather than being an absolute attribute of a test-taker. Therefore, depending on the sample of test-takers that take a particular test, the items' difficulty may vary, and test-takers with different skill levels might receive different results on a simpler or more complex test (Omorogiuwa, 2009).

Comparing test results between different tests is challenging in practice. As a result, IRT, a contemporary theory, was developed to address the shortcomings of CTT (De Boeck & Wilson, 2004; Embretson & Reise, 2010; Nering & Ostini, 2010; Zickar & Broadfoot, 2008). IRT models are advantageous because they tend to fit the data of the 1-, 2-, and 3-parameter logistic models. When applying IRT models to real test data, the fit of the model to the data is the main concern. According to the American Association of Educational Research, American Psychological Association, and National Council on Measurement in Education (2014), obtaining evidence of model-data-fit when using an IRT model to draw conclusions from a data set is the standard for educational and psychological testing. They argue that the adoption of IRT in

estimating genuine data sets is illegitimate if this criterion is not met. Model checking, according to Liu and Maydeu-Olivares (2014), is essential before drawing any conclusions from the fitted model.

Item Response Theory has been utilised in a significant amount of research by both international and Nigerian authors to estimate the item parameters of a scale or test. For example, Wei, Barnard-Brak, Stevens, and William (2018) assessed the item parameter drift of the Self-Description Questionnaire 1 (SDQ1) in relation to children's mathematical self-concept using the IRT paradigm. Their findings indicated that the SDQ1 items exhibit adequate discrimination based on IRT evaluations, suggesting a reexamination of the age appropriateness of the SDQ1.

In addition, Loken and Rulison (2010) investigated the design and justification of a four-parameter IRT model (4-PM) and effectively recovered parameter estimates for respondents and items using a Bayesian technique. Their study concluded that utilising the 4-PM item response model improves overall fit for data created using that model, compared to the 3-PM or 2-PM. They recommended the development of suitable assessment models in psychology and education to better represent the underlying reaction process.

However, Galdin and Laurencelle (2010) used a Monte Carlo analysis to evaluate parameter invariance in the logistic two-item parameter model of IRT. Their findings showed that the θ estimate is inherently biased, and IRT parameters do not outperform or provide additional information beyond those employed in CTT.

Furthermore, Lotobi and Basil (2019) utilised the IRT approach to determine the item parameters for the 2011 basic science test items in the Delta State Basic Education Certificate examination (BECE). They found that 38 items satisfied the combined estimates of the three IRT parameter estimates, while three items (45, 45, and 40) satisfied the IRT difficulty, discrimination, and guessing parameters, respectively.

Shogbesan and Faleye (2021) examined the sensitivity of IRT psychometric estimates to item compromise for the Economics Achievement Test (EAT) administered in secondary schools in Ogun State using one-, two-, and three-parameter logistic models. Their research showed no discernible difference in the item parameter estimates for compromised and secured EAT items across the logistic models. They advised test developers and experts to consider test item security and the sensitivity of IRT parameters when assessing the stability of test items.

To determine if the WAEC and NECO SSCE mathematics multiple-choice test item parameters from 2015 and 2016 past questions satisfied the IRT statistical criterion across school locations, Ethe and Odjegba (2019) compared the test item parameters. Their conclusions revealed that the majority of the test items did not meet the IRT statistical condition, regardless of school location. They recommended that testing organisations ensure proper validation of test items to align with students' abilities. Therefore, the researcher plans to evaluate the model-data fit of the 4IRCTES in secondary schools in Southwestern Nigeria using the multidimensional graded response IRT model in the current study.

Statement of the Problem

For the most part, Nigerian secondary school teachers still lack an advanced understanding of the 4IR. However, a growing number of their students are beginning to grasp the concept. In most cases, secondary school students in Nigeria understand how to use 4IR tools (such as

computers, smartphones, the internet, and so on.) more than their teachers. The 4IR, which supports modern education, requires changes to both basic training and professional development of teachers. Most Nigerian educators lack professional experience with advanced digital tools and how to use them for statistical procedures and data analysis.

Item parameters should be used to highlight the quality of the test instrument, especially when selecting items and evaluating fit in IRT modeling. Selecting the appropriate model that fits the data well requires an efficient evaluation of the model-data fit.

Furthermore, more advanced IRT software must be used to provide a more accurate evaluation of the item parameters, specifically the discrimination power as evaluated in this study, in order for a test maker to achieve the aforementioned goal. Unlike item discrimination indices used in CTT, this software considers responses from all examinees, not just high- and low-scoring groups. The item discrimination parameter typically ranges from 0.0 to 2.0. However, when conducting item parameter analysis with less advanced software, it eliminates items that do not meet this requirement and forces values to be positive. Unfortunately, many test developers do not exercise caution when determining if an item has negative discrimination. In this case, as the examinee's skill level increases, the likelihood of endorsing a correct response would decrease. Consequently, the test's quality would suffer if subject matter experts did not carefully review such items.

Objective of the Study

The primary goal of the research was to analyse the item parameter and model-data fit of the 4IRCTES items to their underlying construct in secondary schools in Southwestern Nigeria. This was done to provide information on the item discrimination indices of Chemistry teachers with high and low levels of teaching effectiveness, as well as the contributions of the items loading on each of the factors of CTES to their underlying construct. The specific objectives of the study were to:

- estimate the discrimination parameter of all the items of 4IRCTES in secondary schools in Southwestern Nigeria.
- assess the model-data fit of the 4IRCTES to multidimensional graded response IRT model in secondary schools in Southwestern Nigeria.

Research Questions

The study aimed to answer the following research questions:

- What is the discrimination parameter of all the items in 4IRCTES in secondary schools in Southwestern Nigeria?
- What is the model-data fit of the 4IRCTES to the multidimensional graded response IRT model in secondary schools in Southwestern Nigeria?

METHOD

Descriptive survey design was utilised in the study to provide a convenient means for participants to express their opinions or share information regarding a specific phenomenon. The survey method was chosen for this study because it offered the researcher further insight into the population, making it easier to identify any issues or concerns that respondents may have had, ultimately aiding in finding or developing solutions to the research problem. The study population consisted of 4,986 Chemistry teachers from all Federal, State, and privately owned

high schools in Osun and Oyo States in Southwestern Nigeria. A sample of thirty-five Chemistry teachers was rated by Chemistry students and department heads.

Sampling, Validation of Fourth Industrial Revolution Chemistry Teachers Effectiveness Scale (4IRCTES), and Data Analysis

The sample was chosen in two phases: first, it was validated in Oyo State, and then it was pilot tested in Osun State. Oyo State established the face and content validity of the 4IRCTES items in stage one. An initial pool of 206 questions from the 4IRCTES instrument, used for data collection, was assessed by four experts in the fields of educational measurement, evaluation, and psychology. The instrument had two components: Section A solicited information on how the items distinguished between teachers with low and high teaching effectiveness, while Section B included details on the dimensionality of 4IRCTES. Twenty-two items were found to be double-barreled, and 88 items did not reflect the true purpose of 4IRCTES, totaling 110 items that were removed from the scale. Ninety-six items made it through the validation phase.

In the second phase, all Chemistry teachers in Osun State's three Federal Government Colleges (13 total), seven State-owned secondary schools, and three privately-owned secondary schools (3 total) were selected using purposive sampling. Twenty-three Chemistry teachers made up the sample in stage two for pilot testing on 4IRCTES.

Exploratory factor analysis (EFA) was used to analyse the data and select items. Thirty-four items that did not meet the factor loadings requirement of 0.5 or above were eliminated, reducing the original 96 items to 62. Reliability analysis using the Cronbach Alpha method on the 62 items yielded a reliability index of 0.93.

Four skilled research assistants were employed to help administer the instrument, following instructions on research ethics and methodology. Data collection for the 4IRCTES's first and second validation rounds took six months. The instrument was administered three times up to the third stage, with a four-month gap between each administration. Questionnaires were coded and analysed, with responses coded as 1 = very poor, 2 = poor, 3 = moderate, and 4 = good.

The principals of the selected secondary schools received an introduction letter from the researcher's head of department at the Department of Educational Foundations and Counseling, Faculty of Education, Obafemi Awolowo University in Ile-Ife, Nigeria. The researcher administered the instrument for the investigation. For the data analysis of research question one, the 62-item 4IRCTES underwent a rigorous item parameter analysis using the Multidimensional Graded Response Model (MGRM) of the Item Response Theory (IRT). The model-data fit analysis for research question two was based on the absolute model-fit, M2 statistic (Maydeu, Olivares, and Joe, 2005; 2006), along with descriptive model-data fit metrics like the Tucker-Lewis Index (TLI), Comparative Fit Index (CFI), and Root Mean Square Error of Approximation (RMSEA).

FINDINGS

Research Question One: What is the discrimination parameter of all the items in 4IRCTES in secondary schools in Southwestern Nigeria? To answer this question, the parameters of the items in 4IRCTES and the factor loading were obtained. In order to achieve this, the fit of the

data to the IRT model was assessed. The item parameters were analysed using the MGRM of the IRT. The calibrated item parameters are presented in Table 1.

Table 1.

Item Parameter Estimates of 4IRCTES

Factor	Item	a	b1	b2	b3	b4
f1	IT1	1.20	-2.63	-1.76	-0.67	0.31
	IT2	1.54	-2.07	-1.20	-0.25	0.75
	IT3	1.48	-2.18	-1.31	-0.24	0.48
	IT7	1.58	-2.17	-1.23	-0.38	0.54
	IT19	1.59	-1.90	-0.99	-0.12	0.64
	IT9	1.69	-2.12	-1.43	-0.42	0.45
	IT17	1.55	-1.57	-0.67	0.14	0.97
	IT32	2.17	-1.63	-0.80	0.02	0.84
	IT33	1.90	-1.64	-0.79	0.00	0.72
	IT31	1.88	-1.57	-0.85	-0.13	0.70
	IT30	1.86	-1.44	-0.80	-0.07	0.84
	IT34	1.69	-1.70	-0.82	0.18	1.12
	IT25	1.67	-2.00	-1.15	-0.10	0.87
	IT28	1.60	-1.63	-0.68	0.19	1.23
	IT29	1.63	-1.89	-1.04	0.08	0.97
	IT12	1.54	-2.31	-1.38	-0.34	0.72
f2	IT20	2.16	-0.88	-0.19	0.71	1.61
	IT22	2.60	-0.61	0.08	0.80	1.49
	IT23	2.00	-0.51	0.22	1.05	1.82
f3	IT48	1.65	-1.56	-0.66	0.22	1.18
	IT49	1.90	-1.45	-0.65	0.24	1.22
	IT50	1.80	-1.57	-0.68	0.30	1.31
	IT51	1.71	-1.85	-0.83	0.27	1.27
	IT52	1.43	-2.12	-1.06	0.19	1.18
	IT53	1.56	-1.96	-0.92	0.23	1.22
f4	IT13	1.70	-1.82	-1.08	-0.12	0.77
	IT14	1.77	-1.76	-1.02	-0.08	0.80
	IT10	1.63	-2.02	-1.42	-0.45	0.52
	IT11	1.75	-2.04	-1.23	-0.35	0.68
f5	IT38	1.96	-1.49	-0.53	0.40	1.35
	IT39	1.47	-1.92	-0.84	0.16	1.26
	IT37	1.62	-1.50	-0.70	0.41	1.51
f6	IT41	1.49	-1.79	-1.01	-0.03	0.82

IT43	1.62	-1.62	-0.79	0.15	1.01
IT42	1.46	-1.77	-0.95	0.22	1.27

Source: Author's Analytical Result

The parameter estimates for the 4IRCTES were presented in Table 1. In Table 1, 'a' represents the item discrimination parameter, while the remaining columns (b1 through b4) display the category boundaries for the items. Each threshold parameter indicates the latent trait level needed to have a 50% or higher likelihood of selecting a specific answer category. All items in Table 1 had discrimination parameters ranging from 1.20 to 2.60, making the 4IRCTES items highly effective in evaluating Chemistry teachers. Moreover, Table 1 illustrated that each item had appropriate boundary placement, with the probability of choosing a response option increasing as one moves between boundaries.

For example, Chemistry teachers with very low effectiveness (-2.63) had a 50% chance of selecting "very poor," while those with low effectiveness (-1.76) had a 50% chance of choosing "poor." Teachers with moderate effectiveness (-0.67) had a 50% chance of selecting "moderate," and teachers with high effectiveness (0.31) had a 50% chance of choosing "good" for item 1 (The Chemistry teacher being assessed encourages students to assist one another). These results indicate that the 4IRCTES items can effectively differentiate between ineffective and effective teachers.

Research Question Two: What is the model-data fit of the 4IRCTES to the multidimensional graded response IRT model in secondary schools in Southwestern Nigeria? In order to address this research question, the model fit analysis employed descriptive model-data fit measures such as the TLI, RMSEA, and CFI in conjunction with absolute model-fit, or M2 statistic (Maydeu, Olivares, and Joe, 2005; 2006). When the p-value for M2 is more than 0.05, the RMSEA is less than or equal to 0.05, and the CFI and TLI are more than or equal to 0.9, the model is considered fit. The results of 4IRCTES's model-data fit are displayed in Table 2.

Table 2.

Model-data Fit of Fourth Industrial Revolution Chemistry Teachers' Effectiveness Scale (4IRCTES)

	M2	df	p	RMSEA	TLI	CFI
stats	1054.755	440	0.062	0.048189	0.916782	0.938131

Source: Field data (2022)

The model-data fit of 4IRCTES to the multidimensional graded response IRT model was displayed in Table 2. According to the Table ($M2(df = 440) = 1054.755$, $p > 0.05$; RMSEA = 0.048; TLI = 0.91; CFI = 0.93), the data fit the model rather well. This implies that 4IRCTES is a good fit, and that MGRM is appropriate for its calibration.

DISCUSSION

This section of the study will discuss the key conclusions drawn from the investigation and, where appropriate, make connections between the research findings and the literature.

Estimation of the discrimination parameter for all items in 4IRCTES

The first research objective involved estimating the factor loading and discrimination parameters for each of the 4IRCTES items, as well as determining the data fitness of the IRT

model. A multidimensional GRM of IRT was used to evaluate the item parameters, and confirmatory full information factor analysis was utilised to analyse the factor loading. This analytical approach aligns with the findings of Edwards and McCullum (2013), who suggested that researchers can proceed to specific IRT models, such as multidimensional IRT models, for confirmatory analysis once they have an understanding of the possible latent dimension(s) of the test data from exploratory approaches. The investigation revealed that the item discrimination indices (a) of the 4IRCTES had high values. This finding supports the suggestion made by Ojerinde, Popoola, Ojo, and Onyeneho (2012) that the values of the a-parameter can vary from - to +, with typical values for items within a psychological construct being less than or more than 2.0. Therefore, the 4IRCTES items were efficient in differentiating between teachers who were less effective and those who were highly effective.

The research findings also support the claims made by Cline and Schwartz (1999), Kaul, Johnsen, Witte, and Saxon (2015), and Seigle, Gubbins, O'Rourke, Langley, Mun, Luria, et al. (2016) that highly effective Chemistry teachers should be able to instruct both highly capable students and those who have little to no motivation or interest in learning to become highly effective in practical Chemistry. Additionally, the study aligns with Sagor and Cox (2013), who found that competent Chemistry teachers can provide at-risk students with the instruction they need to succeed and perform above expectations. The study also found that as the likelihood of selecting any of the response options—very poor, poor, moderate, or good—increases from one boundary to the next, all of the scale's items had relevant boundary positions.

The results of this research are consistent with Ron, Leo, and Steve (2007), who observed that every item has a slope parameter that describes it and that there is a 0.50 likelihood that it falls between category threshold parameters. The results also support the findings of Ethe and Odjegba (2019), who stated that items with low discrimination indexes, nearly zero, are problematic because they fail to distinguish between teachers of differing effectiveness. In general, items with higher 'a' values are preferred over those with lower discrimination values. This implies that the 4IRCTES items were successful in measuring the effectiveness of Chemistry teachers in the 4IR. The study confirms the findings of Wei, Barnard-Brak, Stevens, and William (2018), who found that the SDQI items have adequate discrimination indices based on IRT evaluations.

Assessment of the model-data fit of the 4IRCTES to a multidimensional graded response IRT model in secondary schools in Southwestern Nigeria

The second research objective evaluated how well the Multidimensional Graded Response Model (MGRM) of the IRT fit the 4IRCTES. This study discovered that the multidimensional graded response IRT model significantly fit the 4IRCTES. This is because the RMSEA is less than the criterion (0.05), the CFI and TLI are greater than the basis (0.9), and the p-value of the 4IRCTES linked with M2 is greater than the benchmark of 0.05. Unlike the GRM of the unidimensional IRT model, 4IRCTES is a GRM of the multidimensional IRT. This study's outcome is consistent with that of Shahzad and Mehmood (2019), who demonstrated that the concept of teaching effectiveness is multifaceted because each item on the scale measures a construct and has numerous dimensions. The implication is that student achievement on school-based and standardised assessments of Chemistry is not only attributable to the underlying construct

of “effectiveness of Chemistry teachers in the 4IR.” Consequently, it suggests that there are several dimensions along which the fundamental construct (effectiveness) as proposed by Little, Goe, and Bell (2009) may be measured. These many dimensions encompass additional factors (such as family, friends, school climate, community support, and school resources), which may contribute to teachers’ effectiveness, and determine how well students do on achievement and school-based assessments.

CONCLUSION AND RECOMMENDATIONS

The main focus of this study has been on the model-data fit and item parameter analysis of the 4IRCTES items to their underlying construct in Southwestern Nigerian secondary schools. The study has shown that the multidimensional graded response IRT model significantly fits the 4IRCTES, and that the item discrimination indices (a) of the 4IRCTES had high values. In summary, the 4IRCTES significantly aligns with the MGRM of the IRT; also, the items on the 4IRCTES effectively distinguished between teachers who were ineffective in using technology and those who were effective.

Based on the study’s findings and conclusions, the following suggestions were made:

1. Nigerian government and private school operators should establish technology-friendly teaching and learning environments at the high school level. This would facilitate teachers’ and students’ access to technology-based education.
2. For item parameter analysis, the IRT is advised rather than the Classical Test Theory (CTT). This is because, unlike the item discrimination indices used in CTT, the IRT uses more sophisticated and robust software that provides a more accurate evaluation of the item discrimination power because it considers the responses of all examinees, not just high and low scoring groups.
3. Test developers are advised, with assistance from psychometricians, to carefully examine an item that has a negative discrimination so as to prevent degrading the quality of a scale. This is due to the fact that when an examinee’s ability rises, the likelihood of supporting a valid response should not fall.
4. Teachers should be assigned to classes based on their particular skill levels. This is to ensure that both at-risk and highly capable students are taught by extremely competent teachers. As a result, secondary school teachers will be more equipped to address the particular requirements and traits of various student groups.
5. Education stakeholders, including students, should, when needed, provide sufficient resources to enable successful teaching and learning. This can be achieved by compiling reports and incorporating the actions of various elements that influence students’ learning, such as the home environment, peers, family, school atmosphere, resources, community support, and the work of other teachers rather than just one subject teacher. This is due to the fact that, in contrast to what many parents, policymakers, educators, and other education stakeholders

believe, a student's performance on school-based and/or standardised tests is not solely the responsibility of the teacher. Numerous other factors mentioned above also have a role.

The study may have certain significant shortcomings. For instance, the study's validation of the 4IRCTES only used factor analysis and the IRT's GRM; alternative IRT models, such as the RSM, PCM, GPCM, and NRM, might also be used.

Possible Conflicts of Interest

The author has not disclosed any conflicts of interest.

REFERENCES

Adesoji, F. A. (2008). Students ability levels and effectiveness of problem-solving instructional strategy. *Journal of Social Physics, 17, 5-8.*https://doi.org/10.4236/ce.2019.1012229

Bugaje, B. M. (2013). Qualitative Chemistry education: The role of the teacher. *Journal of Applied Chemistry,* 4(5), 10-14. https://doi.org/10.9790/5736-0451014

Cline, S., & Schwartz, D. (1999). *Diverse populations of gifted children: Meeting their needs in the regular classroom and beyond*. Upper Saddle river, NJ: Merill.

Darling-Hammond, L. (2015). The flat world and education: How America's commitment to equity will determine our future. New York, NY: College Press.

De-Boeck, P., & Wilson, M. (2004). *Explanatory item response model: A generalized linear and nonlinear approach.* New York: Springer.

Edomwonyi, I. O., & Avaa, A. (2011). The challenge of effective teaching of Chemistry: A case study. Accessed 15 July, 2020 from https://www.iejpt.academicdirect.org/A18/00/-008.htm

Edwards, M. C., & McCallum, R. C. (2013). *Current topics in the theory and application of latent variable models* (1st ed.). New York: Routledge Academic.

Embretson, S. E., & Reise, S. P. (2010). *Item response* (2nd ed.). New York: Routledge Academic.

Ethe, N., & Odjegba, O. G. (2019). Comparison of WAEC and NECO SSCE mathematics multiple choice test item parameters across school location: Application of item response theory. *Journal of evaluation,* 4(1), 194-205.

Galdin, M., & Laurencelle, L. (2010). Assessing parameter invariance in item response theory's logistic two-item parameter model: A Monte Carlo investigation. *Tutorials in Quantitative Methods for psychology,* 6(2), 39-51. https://doi.org/10.20982/tqmp.06.2.p039

Hassan L. G., & Salihu, M. (2019). Chemistry education at a crossroad in Nigeria. *Al-Hikmah Journal of Arts and Social Sciences Education*, 1(2), 67-74.

Kaul, C. R., Johnson, S. K., Witte, M. M., & Saxon, T. F. (2015). Critical components of a summer enrichment program for urban low-income gifted students. *Gifted Child Today*, 38(1), 32-40.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology,* 26, 563-575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Little, O., Goe, l., & Bell, C. (2009). A practical guide to evaluating teacher effectiveness. National Comprehensive Center for Teacher Quality, Washington D. C.

Loken, E., & Rulison, K. L. (2010). Estimation of a 4-parameter item response theory model. *The British Journal of Mathematical and Statistical Psychology,* 63(3), 509-525. https://doi:10.1348/000711009X474502, 18 July, 2019.

Lotobi, R. A., & Basil, O. (2019). Determination of item parameters in 2011 basic science test items in Delta State basic education certificate examination using IRT approach. *Nigerian Journal of Educational Research and Evaluation,* 18(1), 71-82.

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and Full-Information Estimation and Goodness-of-Fit Testing in 2n Contingency Tables: A Unified Framework. Journal of the American Statistical Association, 100(471), 1009–1020. https://doi.org/10.1198/016214504000002069

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, 71, 713-732. https://doi.org/10.1007/s11336-005-1295-9

Nering, M. L., & Ostini, R. (2010). *Handbook of polytomous item response theory models.* New York: Routledge.

Holy Bible, New Living Translation (2004). Tyndale House Publishers (original work published 1996).

Nunnally, J. C., & Bernstein, I. A. (1994). *Reliability and validity theory* (3rd ed.). New York: MC.

Ojerinde, D., Popoola, K., Ojo, F., & Onyeneho, P. (2012). Introduction to item response theory: Parameter models, estimation and application (2nd ed.). Abuja, Marvelouse Mike press Ltd.

Omorogiuwa, K. O. (2009). An empirical comparison of the classical test theory and item response theory in the selection of physics achievement test items. A Ph.D. dissertation of faculty of education, University of Benin, Benin city.

Orheruata, M. U. (2015). Item parameter drift of 2012 to 2014 WAEC and NECO SSCE agricultural science multiple-choice items using item response theory. Unpublished Ph.D. dissertation of faculty of education, University of Benin, Benin City.

Hays, R. D., Morales, L. S.., & Reise, S. P. (2000). Item response and health outcomes measurement in the 21st century. *Med care,* 38(9 suppl), 1128-42. https://doi.org/10.1097/00005650-200009002-00007

Sagor, R., & Cox, J. (2013). *At-risk students: Reaching and teaching them* (2nd ed.). New York: Routledge.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 43(4, pt. 2), 100.

Shahzad, S., & Mehmood, N. (2019). Development of teaching effectiveness scale for university teachers. *Journal of research in Social Sciences*, 7(2), 1-14. https://doi.org/10.52015/jrss.7i2.74

Shogbesan, Y. S., & Faleye, B. A. (2021). Sensitivity of economics multiple-choice item parameters to item compromise among secondary school students in Ogun State, Nigeria. *Nigerian Journal of Educational Research and Evaluation,* 20, 267-285.

Siegle, D., Gubbins, E. J., O'Rourke, P., Langley, S. D., Mun, R. U., Luria, S. R. et al. (2016). Barriers to undeserved students' participation in gifted programs and possible solutions. *Journal for the Education of the Gifted,* 39(2), 103-131. https:/doi.org/10.1177/0162353216640930

Stronge, J. H. (2018). *Qualities of effective teacher* (3rd ed.). Alexandria, Virginia: ASCD.

Tabachnik, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Allyn and Bacon, Pearson Education Inc.

VanTassel-Baska, J., & Hubbard, G. F. (2016). Classroom-based strategies for advanced learners in rural settings. *Journal of Advanced Academics*, 27(4), 285-310. https://doi.org/10.1177/1932202X16657645

Wei, T., Barnard-Brak, L., Stevens, T., & William, Y. L. (2018). Item parameter drift of the self-description questionnaire 1: Implications for assessing young children's mathematics self-concept. *European Journal of Psychological Assessment,* 35(6), 1-12. https://doi.org/10.1177/0011000006288127

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The counselling Psychologist,* 34 806-838. https://doi.org/10.1177/0011000006288127

Zickar, M. J., & Broadfoot, A. A. (2009). The partial revival of a dead horse? Comparing classical test theory and item response theory. In C. E. Lance & R. J. Vandenberg (Eds.), Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences (pp. 37–59).