Art. #2356, 15 pages, https://doi.org/10.15700/saje.v45n3a2356

Application of the multidimensional 4-parameter logistic model in the estimation of the psychometric qualities of the West African Senior School Certificate chemistry examination

John J. Agah
Department of Science Education, University of Nigeria, Nsukka, Nigeria
Onisoman Chuks Zudonu
Department of Chemistry Education, Federal College of Education (Technical) Omoku, Rivers State, Nigeria
Basil C. E. Oguguo
Department of Science Education, University of Nigeria, Nsukka, Nigeria
basil.oguguo@unn.edu.ng
Utibe U. James
Uyo Faculty of Education, Akwa Ibom State University, Uyo, Nigeria
Akaneme I. Nwakaego and Francisca C. Okeke
Department of Educational Foundations, Faculty of Education, University of Nigeria, Nsukka
Catherine U. Ene, Samuel Uchenna Nwani and Thaddeus Onyebuchi Ukwueze
Department of Science Education, University of Nigeria, Nsukka, Nigeria

Abstract
In the study reported on here we assessed the dimensionalities and trends in psychometric qualities of the West African Senior School Certificate chemistry examination (WASSCCE) by applying a multidimensional 4-parameter logistic model of item response theory. Trend study was adopted as the design of the study. Four thousand students (n = 4,000) participated in the study and were selected using a multi-stage sampling procedure. The data were collected through the direct delivery technique. Research question 1 was answered using the normal ogive harmonic analysis robust method while questions 2, 3, 4, and 5 were answered using the multidimensional 4-parameter logistic model of item response theory. Microsoft Excel was employed to show trends in the psychometric qualities. The results show that only 1 factor underlies the WASSCCE for 2015 and 2016 while 2 factors underlie WASSCCE for 2017 and 2018. The results shows that WASSCCE for 2015 to 2018 were all multidimensional tests and that the items in WASSCCE 2015 were more difficult than those of 2016, 2017 and 2018. Based on the study we recommend that the psychometric qualities of tests should be determined to understand and adequately compare the quality and performance of the examinees that take the tests.

Keywords: chemistry education; examination; 4-parameter logistic-model; item response theory (IRT); multidimensional; psychometric qualities

Introduction
A test is a set of uniform tasks which examinees respond to individually. The result of such a test is treated in such a way that it provides valid quantitative outcomes. The outcome of tests provides information in the form of scores used to make valid judgments about the test, the examinees, the teachers, the educational system and the quality of the teaching and learning processes (Skidmore, 2017). Tests are applied in virtually every sector of life; they are used for recruitment into public service, for certification, for promotion into a higher office, for offers of admission into institutions of higher learning, for placement purposes, for evaluation of the effectiveness and modification of instructional strategies, to diagnose areas of strength and areas of possible improvement of students, to evaluate educational systems and even business organisations (Jang & Roussos, 2007; Zhang, 2006). One of the primary purposes of testing in the educational system is to provide a means of fairly and objectively measuring or evaluating the ability and skills of a group of examinees (Kolen & Brannen, 2014).

In educational testing, different assessment methods are used to determine the outcome of instruction and to assess the student' ability levels. Test items are developed to reflect the abilities to be assessed by such tests. Chemistry is a core senior secondary school science subject that fosters and consolidates students' knowledge and skills in sciences. Chemistry is an important subject, especially to science students, because it is a major criterium for offer of admission into science and science-related courses at institutions of higher learning. Thus, a credit pass is needed for students' eligibility for tertiary studies in fields such as medicine, engineering and pharmacy, among others (Ogunleye & Babajide, 2019). However, over the years, students' performance in chemistry in high-stakes examinations such as the West African Senior School Certificate examination (WASSCE) has been a source of concern for educators, policymakers, and examination bodies.

Some studies highlight the fluctuating performance trends in chemistry at secondary school level in Nigeria, often attributing the poor outcomes in the subject to factors such as inadequate instructional materials, teacher competency, and student preparedness (Obiekwe & Okoye, 2019; Olatoye & Aderogba, 2019). In a study on the performance of students in chemistry in the WASSCE, Obiekwe and Okoye (2019) revealed that a

significant percentage of students failed to achieve credit passes over a 5-year period. In another study, Olatoye and Aderogba (2019) concluded that schools with well-equipped laboratories delivered better student outcomes. Akpan and Umoren (2018) report that interactive and inquiry-based teaching methods significantly enhance students' understanding and retention of complex concepts. Nwosu and Adesina (2020) report a strong positive correlation between students' attitudes and their performance in high-stakes examinations. Additionally, Adebayo and Fakorede (2021) submit that poorly constructed chemistry examination items often lead to ambiguous interpretations and, consequently, lower performance in the subject. These studies demonstrate the multifaceted nature of factors influencing chemistry performance, ranging from teaching methodologies and resources to the psychometric quality of examination items.

Regarding the psychometric quality of examination items, as a standardised examination, the WASSCE questions pass through rigorous phases, procedures and processes of test development in order to ensure the validity and reliability of the examination questions. According to Umobong and Tommy (2017), the West African Examination Council (WAEC) adopts the classical test theory (CTT) framework other than the novel item response theory (IRT) framework to determine the validity indices (a major component of psychometric qualities) of these examinations. CTT is a framework that describes the relationship between the true score and observed score in a linear fashion, whereas IRT is a group of mathematical models that attempts to explain the relationship between latent traits (unobservable characteristics or attributes) and their manifestations (i.e. observed outcomes, responses or performance). This makes CTT models easy to understand and apply. It is based entirely on total scores or number-of-correct-answer scores. An examinee's observed score is the total score obtained by each examinee and it is different from the true score by a common error score.

The use of CTT in determining the validity indices of examinations has serious deficiencies because of the limitations of the CTT framework, hence, the alternative framework (IRT) corrects most of the limitations. IRT provides procedures for obtaining information on students and assessment items (Zondo, Zewotir & North, 2021). IRT is a collection of mathematical models that uses parameters to characterise both the items in an instrument and the respondents to predict item performance. The relationship between the ability or attribute measured by the instrument and the response to an item is depicted in IRT models.

Two types of IRT models are available: unidimensional models and multidimensional models. Unidimensional models are the unidimensional 1-parameter logistic model (U1PLM) and the unidimensional 2-, 3- and 4-parameter logistic models (U2PLM; U3PLM; U4PLM). The multidimensional models are the following: multidimensional 1-parameter logistic model (M1PLM), and the multidimensional 2-, 3- and 4-parameter logistic models (M2PLM; M3PLM; M4PLM). Due to type of data, the assumptions of IRT, and the data that best suit the model, any of these models could be used (Okwilagwe & Ogunrinde, 2017). IRT rests on the assumption of unidimensionality of the latent traits and local independence (Zondo et al., 2021). Unidimensionality of the latent traits implies that the examination items collectively measure only one latent trait that influences the student scores, with other factors being treated as random errors (DeMars, 2010). The local independence assumption indicates that if the assumption of unidimensionality holds, a student's score in one item will be independent of their score in another item (Zondo et al., 2021).

In educational testing it had been found that cases of multiple proficiencies which are also reflected in the scores obtained for such examination, are present. Proper understanding of the structure of test scores enables an individual to draw meaningful conclusions based on the test scores (Umobong & Tommy, 2017). When the total scores on a test are not from homogeneous items (i.e., when the items measure different ability levels), the interpretation of the total test score may be technically invalid and the meaning of the total test score tends to be ambiguous because the exact measure that the scores represents is not clear (Umobong & Tommy, 2017). For test items to effectively assess what it is designed to measure, the test must be valid. Obtaining the validity indices of each of the test items may be used to establish the validity of the test.

The validity indices of test items can be determined using an item analysis procedure. Item analysis is the process of analysing the quality of test items using statistical approaches. Item analysis (Lee, 2019) is the process of analysing students' responses to distinct examination questions in order to assess the quality of the examination. It is a crucial tool for ensuring test efficacy and impartiality. Donald, Jacobs, Razavieh and Sorensen (2009) define item analysis as the statistical procedure for selecting items for inclusion in a psychological examination. We regard item analysis as a procedure for establishing the psychometric qualities of test items in order to determine an item's ability to distinguish between examinees with high and low abilities, as well as the ability of items to distinguish between competent and less competent students.

Notably, the psychometric qualities of a test determine the quality of items and decision-making

with regard to the scores obtainable and consequently, the interpretation of the scores. The multidimensional nature of chemistry as a subject makes it an ideal subject for psychometric analysis, particularly with models like the 4-parameter logistic model, which can provide insight into the difficulty, discrimination, guessing, and item response functions of examination items (Ezeugo & Onwuka, 2022). For examination bodies such as the West African Examinations Council (WAEC), the insight gained from the psychometric analysis may guide the construction and validation of future examination items, ensuring that they accurately assess students' knowledge and skills. This is particularly important in maintaining the credibility and fairness of high-stakes examinations, which have a direct impact on students' academic and professional trajectories (Abubakar & Usman, 2021). By identifying the specific characteristics of well-functioning items, educators can better align their teaching strategies with the demands of high-stakes examinations, ultimately leading to improved student outcomes (Ogunleye & Babajide, 2019). Thus, since CTT, which is commonly adopted, has serious limitations that may negatively affect decisions and interpretations about scores arising from CTT-based tests, we adopted IRT in assessing the psychometric qualities of the West African Senior School Certificate chemistry examination.

Literature Review

Two major frameworks may be used for item analysis. These are the CTT framework and the IRT framework. The CTT is a framework that describes the relationship between the true score and observed score in a linear fashion. This makes CTT models easy to understand and apply. It is based entirely on total scores or number-of-correct-answer scores. An examinee's observed score is the total score obtained by each examinee and it is different from the true score by a common error score. IRT is a group of mathematical models that attempt to explain the relationship between latent traits (unobservable characteristics or attributes) and their manifestations (i.e. observed outcomes, responses or performance) (De Ayala, 2009). Due to the limitations of the CTT framework, IRT has been recommended as it is more reliable in determining the validity indices of test items (Collins, Chambers & Prather, 2018; De Ayala, 2009). The IRT framework has a number of advantages over the CTT. Item difficulty and item discrimination indices depend on the group in the CTT framework. This indicates that these indices are reliant on the examinees from whom they were acquired, although item characteristics may be obtained within the framework of IRT that is not dependent on a group of examinees. The observed and true test scores are dependent in the CTT framework, meaning that they rise and fall with changes in test difficulty, and the ability estimates are less precise for students of low and high abilities than for students of average ability. Ability scores in the IRT framework do not depend on tests and provide a measure of precision for each ability level. Validity indices such as test dimensionality, item difficulty, item discrimination, and guessing factor indices may be efficiently calculated using the IRT framework.

Dimensionality of the test is one of the major assumptions of the IRT framework. It establishes or strengthens the construct validation of the test. De Ayala (2009) describes dimensionality as the number of traits or constructs assessed by the items in a test. A test assessing only one trait is termed unidimensional, while a test assessing more than one trait of the examinee is referred to as multidimensional (DeMars, 2010). Collins et al. (2018) and Kose and Demirtasli (2012) maintain that when a test tends to measure only one trait, such a test is termed a one-dimensional test but when a test measures two distinctive traits it is termed two-dimensional test, and if a test measures three distinctive traits, it is termed a three-dimensional test.

Assessing dimensionality of test items is very important in the interpretation of scores obtained from that test. In circumstances where the items in a test are not homogeneous (testing same distinctive constructs), the interpretation of that test based on one construct using the scores produced from such a test will be substantially insufficient. It is possible that the interpretation of the test score as indicating the construct that the test is supposed to measure is incorrect. The overall score would include this information if an item captured not only the desired construct but also other constructs. The test result could no longer be inferred as a person's position on the construct. As a result, the dimensionality of the items that make up a test score is critical for validity of the evaluation processes that the score is meant for. An interpretation of the test score as indicating one dimension without checking for dimensionality is possibly dangerous (Ziegler & Hagemann, 2015). The dimensionalities must be tested in order for the results of the West African Senior School Certificate chemistry examination to be properly reported, understood, and compared. This helps to ensure the fairness, validity and reliability of chemistry test items. It also helps to ensure that chemistry students are not subjected to tests with items that are beyond their ability levels which will likely lead to a high failure rate in such examinations.

Various methods exist for assessing the dimensionalities of a test. The first is the principal component factor analysis, which is often used for

polytomously scored data and is included in the Statistical Package for Social Sciences (SPSS). Another is Stout's test of essential unidimensionality, which is commonly referred to as DIMTEST because it is calculated using the DIMTEST 2.0 software (Stout, 2005). This approach, however, has the limitation of only providing information that a test has either one dimension or more than one dimension. Put differently, it cannot provide information of how many dimensions underly a test if the test has more than one dimension. The second is the use of residuals of the bivariate proportion-correct matrix from a unidimensional IRT model. The residual bivariate proportion-correct matrix is provided by one IRT programme; NORHAM 3 (Fraser & McDonald, 2003). NORHAM has an advantage over the DIMTEST in that it offers information on the actual number of dimensions that a test is based on. As a result of this explanation, we used NORHARM to examine the dimensionalities of the 2015 to 2018 chemistry examinations. Assessing the dimensionality of the chemistry examination ensures that test items are fair, reliable and valid since the scores obtained from such examinations are expected to reflect the students' true abilities.

By comparing the root mean square (RMS) value obtained from the residual matrix, which is the square root of the average square difference between the observed and predicted covariance, with the RMS criterion value, which is four times the reciprocal of the square root of the sample size, the number of dimensions that generate a better model can be verified using NORHAM to test the dimensionality of a test. Unidimensionality is defined as a percentage root mean square reduction (RMSR) of less than 10 (R per cent 10) whereas multidimensionality is defined as a percentage reduction equal to or more than 10 (R per cent 10).

The parameter that determines how the item behaves along the ability scale is item difficulty (b). It is a measure of the percentage of examinees that correctly answered the item (Maydeu-Olivares, 2015). Determined at the median point of likelihood, according to Hingorjo and Jaleel (2012), it is the ability at which 50% of respondents provide the correct answer. On an item characteristic curve, items that are harder to answer are shifted to the right of the scale, indicating that respondents who correctly answer those have a higher ability, whereas items that are easier are shifted to the left of the ability scale. It is also used to express how difficult it is to have a 0.5 chance of getting a right response for an item based on the respondent's latent variable level (Miller, Linn & Gronlund, 2009). The greater the ability level required reaching this aim, the more difficult it is for a student to have a 50% chance of successfully answering an item (Obinne, 2012).

The ability to identify high- and low-scoring students is measured by the discrimination index, (a), of an item. The closer this number is to 1, the better the item distinguishes high-scoring students from low-scoring students (Obinne, 2012). By identifying whether or not the items are working properly, the discrimination index enhances validity and reliability of tests. It calculates the likelihood of correctly answering an item as ability level changes. It is critical in distinguishing examinees with identical degrees of latent construct of interest (Maydeu-Olivares, 2015). According to Ziegler and Hagemann (2015), the ultimate goal of creating a precise measure is to include items with high discrimination so that individuals may be mapped along the latent trait continuum. The item discrimination index is a measure of how successfully an item can discriminate between knowledgeable and non-knowledgeable examinees (Miller et al., 2009).

The analysis of each item, which includes calculating difficulty and discrimination indices, offers feedback on what the students have learned and allows teachers to identify and fix those items that are faulty. The failure to do item analysis to identify the difficulty indices and discrimination indices of chemistry tests has the consequence that the majority of the items may be faulty, particularly in classifying examinees based on ability. Item difficulty and discrimination play a role in distinguishing examinees of average ability and classifying examinees of various capacities (Hambleton & Swaminathan, 1981; Sorum, 1958). According to the most basic form of IRT, the likelihood that the student will respond correctly to a particular test item is affected by two things: the student's ability and the item's difficulty. Therefore, the probability that a particular student will respond correctly to a given test item depends on his/her ability and the level of difficulty of the test item. Just as the test items differ in terms of their level of difficulty, they might also differ in terms of the degree to which they can differentiate between students with high and low ability levels, which is called item discrimination (Zondo et al., 2021:3). The guessing factor index is another index to evaluate.

Judging something without having complete information about it is no more than guessing. According to Obinne (2012), guessing is a standard test-taking approach offered to examinees taking a multiple-choice examination. This strategy allows the examinee to have an item counted as correct even if they have insufficient understanding of the subject area. If test scores are only determined by the number of questions properly answered, a random guess enhances the likelihood of a better score. According to Delavar and Zahrakar (2013), examinees may use guessing as a tactic to gain extra marks. The examinee may be able to guess an

item based on how it is constructed. Many students and teachers believe that guessing is a major element in determining an examinee's score on an objective examination.

The guessing factor index has serious assessment implications because the more the items in the examination are prone to guessing, the more difficult it becomes to determine the examinees' true abilities. One major effect of guessing is that it has the capacity to affect the psychometric properties of the test. Guessed responses raise the variance error of test scores and reduce the reliability of the test (Bereby-Meyer, Meyer & Flascher, 2002). Guessing increases the measurement error and probable answers, as well as the structural error variance, which is a severe threat to construct validity. Guessing also leads to errors and reduces the relationship between the replies (Yeh, 2007). To avoid these errors and to increase the reliability of a test, the guessing factor index of the items in an examination should be determined.

Careless errors, as compared to guesses, may result in more substantial estimation biases, especially if they occur early in a test. The carelessness parameter is a term used to describe this type of error, indicated by *d*. It was created as a result of the efforts of Barton and Lord (1981). It represents an upper asymptote that does not result in 100% performance for high-ability students (high-ability students fail items that are below their abilities) due to stress or carelessness. The existence of sloppy responses would negatively bias items' difficulty parameter, making the items appear more difficult than they are. This effect may be especially noticeable in speed tests, where the strain of time can lead to careless errors (Boughton & Yamamoto, 2007; Mroch, Bolt & Wollack, 2005; Van der Linden, 2007). According to Magis (2013), Magis and Raîche (2012), and Maniaci and Rogge (2014), the appropriate range of the carelessness parameter is 0.75 to 1.00.

Science students sit for the West African Senior School Certificate chemistry examination (WASSCCE) at the end of their secondary school education to assess the level at which these scientific skills and attributes have been inculcated and developed in the child. The scores obtained from such examination are interpreted and used to make comparisons and decisions about the child, the teachers, the educational system and so on. Such decisions are based on the scores obtained by the students in these examinations which may be faulty and inappropriate if the psychometric qualities of these examinations are not appropriately determined. Most educational test instrument measures multiple proficiencies and as such, these proficiencies also reflect in the total score obtained from the test. So, when such test is compared or decisions are made based on just a hypothesised ability and neglecting other abilities measured by the test, the decision becomes faulty and inappropriate. Hence, the need to determine the psychometric qualities of WASSCCE ensures that the scores obtained by chemistry students in such examinations reflect their true abilities.

Okwilagwe and Ogunrinde (2017) used DIMTEST to investigate the unidimensionality of the WAEC and the National Examination Council's (NECO) 2013 geography achievement tests. The research reveals that the variance in examinee responses to the geography test items can be explained by more than one dimension. Chikezie (2017) used principal component analysis and IRT to analyse the unidimensionality of the WASSCE in chemistry. According to the analysis, the WASSCE chemistry test was not unidimensional. According to Olonade, Metibemu and Adewale (2017), the 2014 WASSCE mathematics multiple-choice test contained four dimensions that best explained examinee performance.

Ene (2014) has documented the item parameters of self-developed multiple-choice items in basic science. The author report that 20% of the items were very easy while 80% of the items were moderately difficult; 87% of the items had moderate discrimination and 13% had low discrimination indices. Ani (2014) also reports on the item parameters of multiple-choice items in economics. The report shows that 66% of the items were easy while 34% of the items were difficult, 20%, 36%, 40% and 4% had very low, low, high and very high discrimination indices respectively, and 90% of the items had good discrimination indices. Zondo et al. (2021) carried out a study to determine the level of difficulty and discrimination power of the items in the National Senior Certificate mathematics examination in South Africa and discovered that the discrimination power of the different examination questions was not identical across different school quintiles. Adonu (2014) reports that the item difficulty estimates for the WAEC and the NECO examinations for 2011 and 2012 showed that all the items had difficulty estimates that ranged between -1.53 to +1.94, which shows that their difficulty was moderate for all items. From 2016 to 2018, Tommy and Udo (2019) found that item difficulty parameters for the NECO biology objective questions were suitable; however they were not grouped hierarchically from least to most difficult. The authors discovered that between 2009 and 2012, there was a 3.30% rise for candidates of the National Examination Council's Senior Secondary Certificate examination (NECOSSCE) compared to candidates for the WASSCE, with 5,940 and 6,136 candidates registered for the WASSCE and NECOSSCE, respectively. In all the reviews, none of the studies adopted the multidimensional

parameter logistic model of IRT. The application of the multidimensional 4-parameter logistic model, for example, offers a robust approach to understanding the various factors that influence item performance beyond the traditional one-dimensional models (Ezeugo & Onwuka, 2022). Also, none of the studies examined trends in test item parameters and none of the studies was carried out comparing the psychometric properties of the WASSCCE for 4 years simultaneously and in the Enugu State. The study discussed in this article resulted from the background stated above.

The purpose of this study was to determine the application of multidimensional 4-parameter logistic model in the estimation of the psychometric qualities of the WASSCCE. The following research questions were addressed:

1) What are the dimensionalities of the West African Senior School Certificate chemistry examination from 2015 to 2018?
2) What are the trends of item difficulty indices of the West African Senior School Certificate chemistry examination from 2015 to 2018?
3) What are the trends of item discrimination indices of the West African Senior School Certificate chemistry examination from 2015 to 2018?
4) What are the trends of the guessing factor indices of the West African Senior School Certificate chemistry examination from 2015 to 2018?
5) What are the trends of carelessness indices of the West African Senior School Certificate chemistry examination from 2015 to 2018?

Materials and Method

We adopted trend study as the design for the study. According to Rae (2014), a trend study is a sort of longitudinal survey research methodology used to collect data and find patterns, or trends, within that data in order to understand or anticipate behaviour. The design was ideal for this study as we were interested in collecting data and assessing the psychometric qualities of the WASSCCE from 2015 to 2018 with a view to identifying the trends. The study was carried out with senior secondary (SS3) science students. The population of science students in the Enugu State for the 2019/2020 academic session was 9,610. We adopted the multi-stage sampling procedure (MSSP). The procedure, according to Ali (2006), requires several stages of sampling elements of a population. We drew a total sample of 4,000 students from the entire population of SS3 students in the Enugu State, Nigeria. We considered this sample size to be appropriate because Sinharay and Lu (2008) recommend that the use of a sample size of 2,000 and more with three-parameter logistic (3PL) helps to control type I error rates much better than most polytomous models. De Ayala (2009) also suggests that a sample of 2,000 and above will, under normal conditions, lead to reasonably accurate item parameter estimates with the 3-PL model.

The instrument for data collection was adopted from the WASSCCE for 2015 to 2018. Four response options, A, B, C, and D, were available for each item, with only one of the options being correct (key). We administered the instrument to the students in the selected schools through the direct delivery technique (DDT) with the help of the chemistry teachers at the schools and 10 research assistants to whom the modalities of the exams were explained. The 10 research assistants, who were also postgraduate students of the University of Nigeria, Nsukka, helped with the administration of the instrument and supervised the conduct of the examination. One hour had been allotted for responding to the instrument. This duration was chosen since the multiple-choice segment of the WASSCCE is allotted the same amount of time. The data gathered through the instrument were examined in relation to the research questions and hypotheses of the study. The dimensionality of the items in question 1 was determined using the normal ogive harmonic analysis robust method (NOHARM) in which the RMSR percentage of less than 10 ($R\% < 10$) is considered to be unidimensional whereas a percentage reduction equal to or greater than 10 ($R\% \geq 10$) is considered to be multidimensional. Research questions 2, 3, 4, and 5 were answered using the multidimensional four parameter logistic model of item response theory (M4PLMIRT) using the R package. This is because the Akaike information criterion (AIC) of each model was compared, and 4-PLM was the model that best fit the data. The trend was compared using Microsoft Excel software.

Ethical Approval and Consent to Participate

The ethical standards set by the institutional research committee of the University of Nigeria, Nsukka, and the 1964 Helsinki declaration were followed while collecting data from the entire sample. Before the start of the study, the respondents were required to complete and sign informed consent forms, which they did.

Results

Table 1 Dimensionalities of the West African Senior School Certificate chemistry examination (WASSCCE) for 2015 to 2018

Year	No. of dim	RMSR	Tanaka's GFI	RMS criterion	SSR	% reduction	Remarks
2015	3	0.0231468	0.7703578	0.063	0.6563217	R > 10	Multidimensional
2016	3	0.0124639	0.9598164	0.063	0.1903008	R = 10	Multidimensional
2017	2	0.0115769	0.9652036	0.063	0.1641812	R > 10	Multidimensional
2018	2	0.0099498	0.9765501	0.063	0.1212740	R > 10	Multidimensional

Note. No. = Number; dim = Dimension; GFI = Goodness-of-fit index; RMSR = Root mean square residual; SSR = Sum of square residual.

The results in Table 1 show the dimensionality of the WASSCCE for 2015 to 2018. The RMSR, GFI, RMS criterion and SSR values for the 2015 WASSCCE were 0.0231468, 0.7703578, 0.063 and 0.6563217 respectively. The RMSR, GFI, RMS criterion and SSR values for the 2016 WASSCCE were 0.0124639, 0.9598164, 0.063 and 0.1903008 respectively. The RMSR, GFI, RMS criterion and SSR values for the 2017 WASSCCE were 0.0115769, 0.9652036, 0.063 and 0.1641812 respectively, and the RMSR, GFI, RMS criterion and SSR values for the WASSCCE 2018 were 0.0099498, 0.9765501, 0.063 and 0.1212740 respectively. The results in the table also shows that the percentage reduction for 2015, 2017 and 2018 were all greater than 10 while the percentage reduction of 2016 was equal to 10. This implies that all the WASSCCEs were multidimensional. Furthermore, three dimensions underlay the WASSCCE for 2015 and 2016 while two dimensions underlay each of the WASSCCEs for 2017 and 2018.

Table 2 Difficulty indices of the West African Senior School Certificate chemistry examination (WASSCCE) for 2015 to 2018

Year	Below acceptable range	Within acceptable range	Above acceptable range
2015	24 (48%) {1, 2, 3, 4, 6, 8, 9, 12, 14, 18, 19, 20, 21, 22, 29, 33, 34, 36, 38, 45, 47, & 50}	8 (16%) {5, 17, 27, 31, 41, 43, 48, & 49}	18 (36%) {7, 10, 11, 13, 16, 23, 24, 25, 26, 28, 30, 32, 35, 37, 40, 42, 44 & 46}
2016	1 (2%) {12}	45 (90%) {1, 4, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, & 50}	4 (8%) {2, 3, 5 & 36}
2017	3 (6%) {12, 15 & 37}	44 (88%) {1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 13, 14, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 38, 39, 40, 41, 42, 43, 44, 45, 46, 48, 49 & 50}	3 (6%) {5, 36 & 47}
2018	Nil (0%)	49 (98%) {1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, & 50}	1 (2%) {2}

Table 2 shows the trends in the item difficulty indices of the WASSCCE for 2015 to 2018. The results in Table 2 show that in 2015, the difficulty indices ranged from -1.68 to 2.78. Only 24 (48%) of the items were below the acceptable range, eight (16%) of the items were within the acceptable range and 18 (36%) of the items were above the acceptable range. In 2016, the difficulty indices ranged from -2.18 to 2.48; only one (2%) of the items was below the acceptable range, 45 (90%) of the items were within the acceptable range and four (8%) of the items were above the acceptable range. In 2017, difficulty indices ranged from -2.36 to 1.80; only three (6%) of the items were below the acceptable range, 44 (88%) of the items were within the acceptable range and three (6%) of the items were above the acceptable range. In 2018, the difficulty indices ranged from -2.20 to 2.73. None of the items were below the acceptable range, 49 (98%) of the items were within the acceptable range and only one (2%) of the items was above the acceptable range.

Figure 1 below shows the trend in item difficulty of the WASSCCE from 2015 to 2018. The graph shows that most of the items in the 2016, 2017 and 2018 WASSCCE were moderately

difficult as all were within the numeric value of zero (at the centre of the graph), whereas the highest level of difficulty for the 2015 WASSCE was at about +100, while the lowest level of difficulty was at almost -150. This is an indication that the 2015 WASSCCE seemed to have more difficult items than other WASSCCEs.

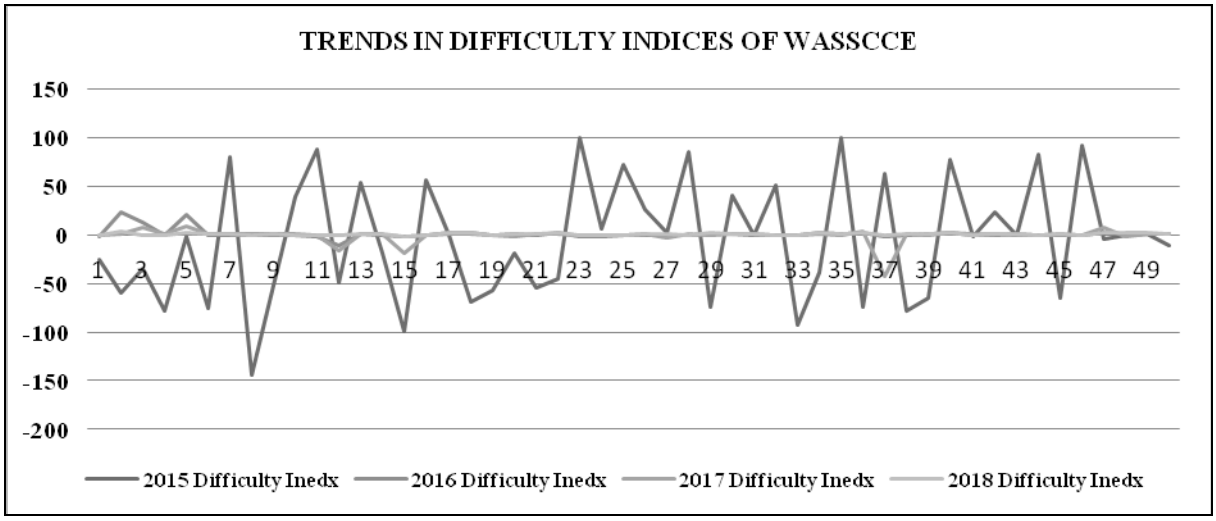

Figure 1 Trends in difficulty indices of the West African Senior School Certificate chemistry examination (WASSCCE) for 2015 to 2018

Table 3 Discrimination indices of the West African Senior School Certificate chemistry examination (WASSCCE) for 2015 to 2018

Year		Below acceptable range	Within acceptable range	Above acceptable range
2015	a1	13 (26%) {3, 16, 19, 20, 22, 23, 25, 26, 28, 30, 35, 42 & 50}	6 (12%) {5, 13, 31, 43, 48 & 49}	31 (62%) {1, 2, 4, 6, 7, 8, 9, 10, 11, 12, 14, 15, 17, 18, 21, 24, 27, 29, 32, 33, 34, 36, 37, 38, 39, 40, 41, 43, 44, 45, 46, 47 & 50}
	a2	27 (54%) {2, 4, 6, 7, 11, 13, 14, 16, 19, 20, 22, 23, 26, 27, 28, 30, 32, 34, 37, 39, 40, 41, 42, 44, 46, 48 & 50}	9 (18%) {5, 17, 25, 29, 31, 36, 43, 47 & 49}	14 (28%) {1, 3, 8, 9, 10, 12, 15, 18, 21, 24, 33, 35, 38 & 45}
2016	a1	13 (26%) {1, 5, 6, 9, 12, 17, 27, 29, 31, 36, 40, 47 & 48}	36 (72%) {2, 4, 7, 8, 10, 11, 13, 14, 15, 16, 18, 19, 20, 21, 22, 23, 24, 25, 26, 28, 30, 32, 33, 34, 35, 37, 38, 39, 41, 42, 43, 44, 45, 46, 49 & 50}	1 (2%) {3}
	a2	3 (6%) {2, 3 & 5}	46 (92%) {1, 4, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, & 50}	1 (2%) {12}
2017	a1	8 (16%) {3, 7, 12, 15, 20, 27, 36 & 37}	41 (82%) {1, 2, 4, 5, 6, 8, 9, 10, 11, 13, 14, 16, 17, 18, 19, 21, 22, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 34, 35, 38, 39, 40, 41, 42, 43, 44, 45, 46, 48, 49 & 50}	1 (2%) {47}
	a2	4 (8%) {2, 3, 5 & 27}	42 (84%) {1, 4, 6, 7, 8, 9, 10, 11, 13, 14, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 34, 35, 36, 38, 39, 40, 41, 42, 43, 44, 45, 46, 48, 49 & 50}	4 (8%) {12, 15, 37 & 47}
2018	a1	9 (18%) {6, 7, 9, 17, 27, 29, 36, 47 & 48}	41 (82%) {1, 2, 3, 4, 5, 8, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 22, 23, 24, 25, 26, 28, 30, 31, 32, 33, 34, 35, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 49 & 50}	Nil (0%) {Nil}
	a2	Nil (0%) {Nil}	48 (96%) {1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, & 50}	2 (4%) {2 & 20}

Table 3 shows the trends in the item discrimination indices of the WASSCCE for 2015 to 2018. The table shows that each of the WASSCCE had two discrimination indices. In 2015 (a1) ranged from -1.96 to +2.87; 13 (26%) of the items were below the acceptable range, six (12%) of the items were within the acceptable range and 31 (62%) of the items were above the acceptable range. In 2015 (a2) ranged from -2.58 to +2.16; 27 (54%) of the items were below the acceptable range, nine (18%) of the items were within the acceptable range and 14 (28%) of the items were above the acceptable range. In 2016 (a1) ranged from -2.78 to +1.44; 13 (26%) of the items were below the acceptable range, 36 (72%) of the items were within the acceptable range and one (2%) of the items was above the acceptable range. In 2016 (a2) ranged from -2.61 to +1.92; only three (6%) of the items were below the acceptable range, 46 (92%) of the items were within the acceptable range and one (2%) of the items was above the acceptable range. In 2017 (a1) ranged from -2.49 to +0.79; only eight (16%) of the items were below the acceptable range, 41 (82%) of the items were within the acceptable range and one (2%) of the items was above the acceptable range. In 2017 (a2) ranged from -2.84 to +0.81; only four (8%) of the items were below the acceptable range, 42 (84%) of the items were within the acceptable range and four (8%) of the items were above the acceptable range. In 2018 (a1) ranged from -2.78 to +1.63; only nine (18%)

of the items were below the acceptable range, 41 (82%) of the items were within the acceptable range and none (0%) of the items were above the acceptable range. In 2018 (a2) ranged from -2.05 to +2.55, none (0%) of the items were below the acceptable range, 48 (96%) of the items were within the acceptable range and two (4%) of the items were above the acceptable range.

Figure 2 below shows the trends in item discrimination indices of the WASSCCE for 2015 to 2018. The results show that the WASSCCE for 2015 had a higher discriminating power than other WASSCCEs. High- and low-ability examinees were more effectively identified in the 2015 WASSCCE than in other examinations, with its highest discriminating point a little above 100 and the lowest discriminating point a little below -150.

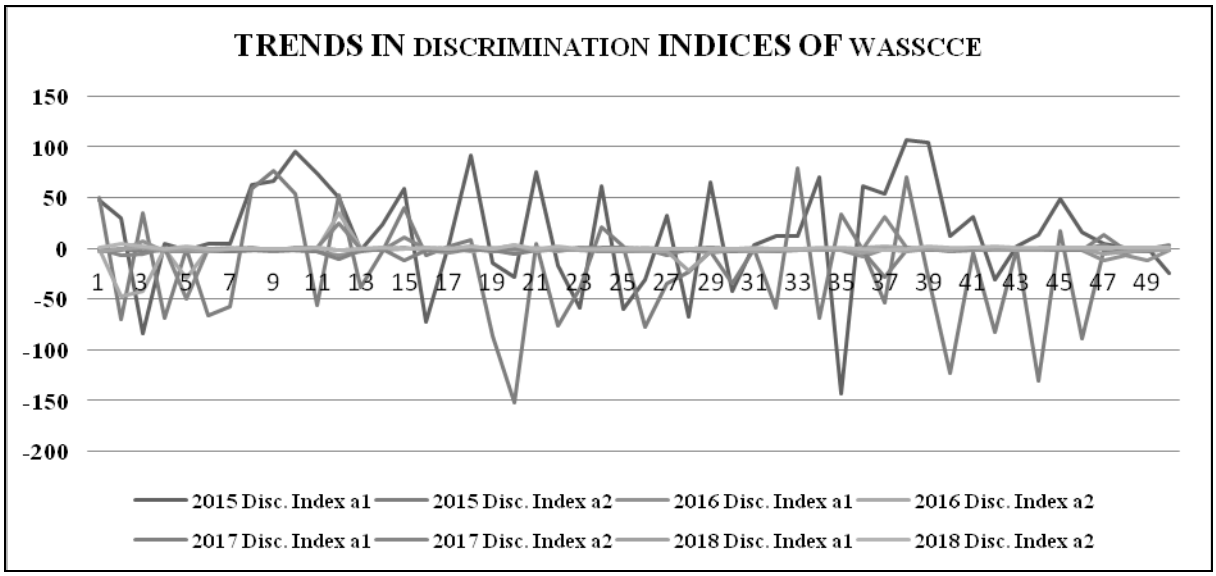

Figure 2 Trends in discrimination indices of the West African Senior School Certificate chemistry examination (WASSCCE) for 2015 to 2018

Table 4 Guessing indices of the West African Senior School Certificate chemistry examination (WASSCCE) for 2015 to 2018

Year	Below acceptable range	Within acceptable range	Above acceptable range
2015	Nil (0%)	50 (100%) {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, & 50}	Nil (0%)
2016	Nil (0%)	50 (100%) {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, & 50}	Nil (0%)
2017	Nil (0%)	49 (98%) {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, & 50}	1(2%) {30}
2018	Nil (0%)	50 (100%) {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, & 50}	Nil (0%)

Table 4 shows the guessing indices of the WASSCCE for 2015 to 2018. The guessing factor indices for the 2015, 2016, 2017 and 2018 WASSCCE ranged from 0.00 to +0.74, 0.00 to

+0.43, 0.00 to +0.51 and 0.00 to 0.42 respectively. From the table, it is clear that all the items in the WASSCCE for 2015 to 2018 were within the acceptable range. However, only one item (Item 30) in WASSCCE 2017 had a guessing index above the acceptable range.

Figure 3 shows the trends in guessing indices of WASSCCE for 2015 to 2018. From the graph it is clear that the items in WASSCCE 2015 were more prone to guessing than others. The graph also shows that one item (Item 30) in the 2017 WASSCCE was highly vulnerable to guessing, making it an inappropriate item with regard to the guessing parameter.

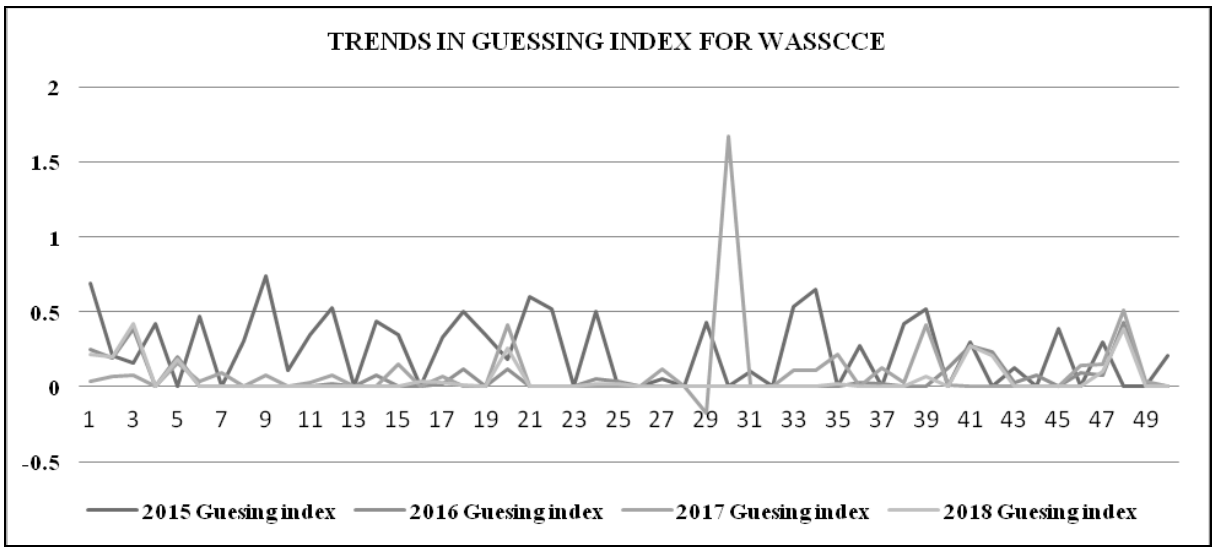

Figure 3 Trends in guessing indices of the West African Senior School Certificate chemistry examination (WASSCCE) for 2015 to 2018

Table 5 Carelessness indices of the West African Senior School Certificate chemistry examination (WASSCCE) for 2015 to 2018

Year	Below acceptable range	Within acceptable range	Above acceptable range
2015	16 (32%) {7, 13, 16, 23, 25, 26, 28, 29, 30, 32, 35, 37, 40, 42, 46, & 50}	34 (68%) {1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 14, 15, 17, 18, 19, 20, 21, 22, 24, 27, 31, 33, 34, 36, 38, 39, 41, 43, 44, 45, 47, 48, & 49}	Nil (0%)
2016	5 (10%) {12, 29, 36, 37, & 47}	45 (90%) {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 31, 32, 33, 34, 35, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 48, 49, & 50}	Nil (0%)
2017	2 (4%) {35 & 46}	48 (96%) {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 47, 48, 49, & 50}	Nil (0%)
2018	6 (12%) {12, 29, 36, 37, 43, & 47}	44 (88%) {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 31, 32, 33, 34, 35, 38, 39, 40, 41, 42, 44, 45, 46, 48, 49, & 50}	Nil (0%)

Table 5 above shows the trends in the carelessness indices of the WASSCCE for 2015 to 2018. The results reveal that in 2015, the carelessness indices ranged from 0.00 to 1.00; 16 (32%) of the items were below the acceptable range, 34 (68%) of the items were within the acceptable range and no item was above the acceptable range. In 2016, the carelessness indices ranged from 0.24 to 1.00; only five (10%) of the items were below the acceptable range, 45 (90%)

of the items were within the acceptable range and no item was above the acceptable range. In 2017, the carelessness indices ranged from 0.50 to 1.00; only two (4%) of the items were below the acceptable range, 48 (96%) of the items were within the acceptable range and no item was above the acceptable range. In 2018, the difficulty indices ranged from 0.26 to 1.00; only six (12%) of the items were below the acceptable range, 44 (88%) of the items were within the acceptable range and no item was above the acceptable range.

Figure 4 shows the trends in the carelessness indices of the WASSCCE for 2015 to 2018. The graph shows that the carelessness index for the 2015 WASSCCE touched the zero (0) line indicating that the bright students were more careless in choosing the wrong answers, which they should have responded to correctly. The carelessness index of the 2018 WASSCCE also shows that most bright students also responded incorrectly to the options to which they should have responded correctly.

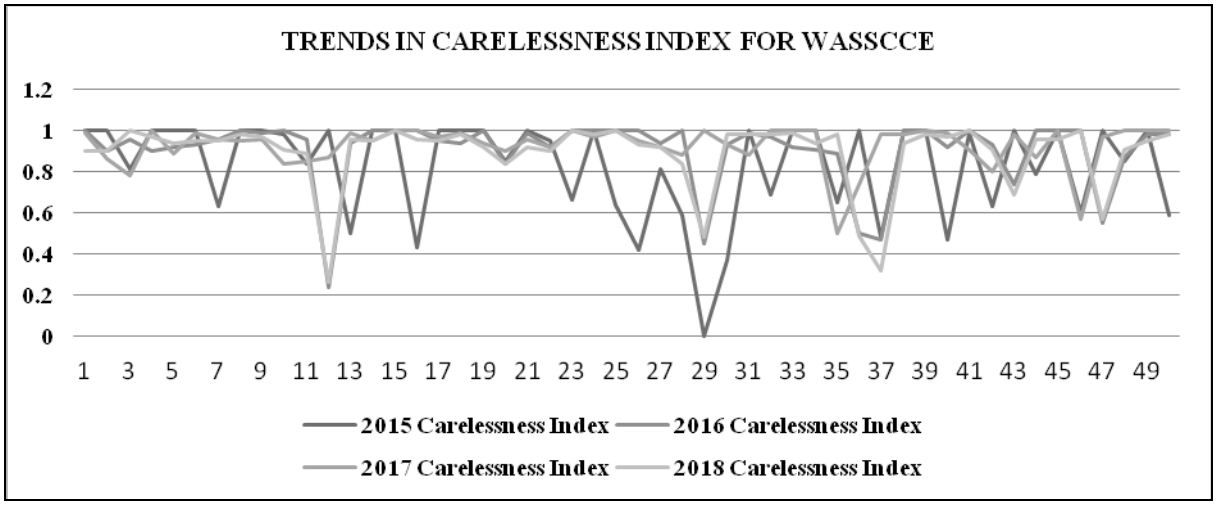

Figure 4 Trends in carelessness indices of the West African Senior School Certificate chemistry examination (WASSCCE) for 2015 to 2018

Discussion

In this study we tested for the dimensionalities of the WASSCCE from 2015 to 2018. The results show that three dimensions accounted for the variation in students' responses to the 2015 and 2016 WASSCCE while two dimensions accounted for the variation in students' responses to the 2017 and 2018 WASSCCE. The results also indicate that the WASSCCE for 2015, 2016, 2017 and 2018 were all multidimensional tests. When a test is multidimensional, it implies that the test measures multiple abilities or construct. In the WASSCCE for 2015, 2016, 2017 and 2018, the tests were multidimensional because the tests had been drawn from the various components of the chemistry curriculum or scheme. It could also be due to the fact that the tests were designed for different cultural groups with diverse ideologies, and to enhance the relevance and its validity across the diverse population. The result agrees with the findings of Okwilagwe and Ogunrinde (2017) that more than one dimension accounted for the variation in examinees' responses to the geography test items. Hence, it was a multidimensional test. The finding is also in line with Chikezie (2017) that the WASSCE chemistry test was not unidimensional. The finding is also in agreement with that of Olonade et al. (2017) that the 2014 WASSCE mathematics multiple-choice test optimally had four dimensions underlying the performance of examinees in the test.

Item parameters of the WASSCCE were obtained for 2015 to 2018. The result shows that the items in the 2016, 2017 and 2018 WASSCCE were moderately difficult, moderately differentiated between high-ability and low-ability students, had a moderate level of vulnerability to guessing and a moderate level of carelessness. This finding could be due to the nature of standardised testing. The moderate levels observed across difficulty, discrimination, guessing, and carelessness suggest that the WASSCCE examination items were constructed well. It also shows that the items were appropriately challenging and distinguished between different levels of student ability while accounting for typical test-taking behaviour. To a large extent, this finding is in line with that in a study by Zondo et al. (2021) who report a considerable range of category difficulty levels, with higher (above average) ability levels being tested for learners in quintile 1 to quintile 4 schools, while only learners with average abilities were being tested in quintile 5 and independent schools.

However, easier items in the 2015 WASSCCE highly differentiated between

high-ability and low-ability students but had a moderate level of vulnerability to guessing as well as a moderate level of carelessness. This is an indication that the items that make up a test possess peculiar psychometric qualities that are different from one another. This could be due to the fact that the chemistry examinations must have covered a wide range of topics or content areas leading to development of items that uniquely assesses those content areas. It could also be that the process of test item development including item writing, item review and validation varies across examinations; the quality and rigor of this process could have affected the uniqueness of the psychometric properties of the items. The finding is in agreement with that of Ene (2014) that only 20% of the items of biology achievement tests were very easy while 80% of the items were moderately difficult, 87% of the items had moderate discrimination and 13% had low discrimination indices. The finding is also in line with that of Ani (2014) that in multiple-choice items in economics, 66% of the items were easy, 34% of the items were difficult, and 20%, 36%, 40% and 4% had very low, low, high and very high discrimination indices respectively. Ninety per cent of the items had good discrimination indices. The findings also agree with the report by Zondo et al. (2021) that the discrimination power of the different examination questions was not identical across different school quintiles in South Africa. The finding also supports that of Adonu (2014) that the item difficulty estimates for the 2011 and 2012 WAEC and NECO examinations showed that the estimates for all the items were moderate difficult.

With this study we have established that every item in a test measures at least one underlying construct. This will enable test developers (examination bodies) to develop and carefully select test items that have the same properties as others included in such an examination. In other words, they should ensure concurrency in test quality (creating a population of questions with known properties, e.g. a test bank). Considering the psychometric properties of test items before administering these to examinees or selecting these as test items is a step to having reliable and valid test items.

We understand that the psychometric quality of a particular examination significantly differs from that of another. This is because the testing of the dimensionalities of the West African Senior School chemistry examination showed the number of dimensions underlying the examination. The validity indices were also not the same. The tests also showed that making decisions, comparisons and judgments based on the scores obtained from these examinations without testing and comparing the psychometric qualities is faulty and inappropriate.

Conclusion

Three factors accounted for the variations in the student responses to the 2015 and 2016 WASSCCE while two factors accounted for the variations in the student responses to the 2017 and 2018 WASSCCE. Hence, the WASSCCE for 2015 to 2018 are multidimensional tests. Since most of the item parameters (item difficulty, item discrimination, guessing factor and carelessness) of the WASSCCE for 2015 to 2018 were within the acceptable range, and thus regarded as good and reliable items. The findings of this study contribute to knowledge by revealing that the IRT demonstrates a thorough understanding of relevant literature. With respect to this study, the theory provides the opportunity to test the dimensionalities and trends in psychometric qualities of WASSCCE. It is most efficient in establishing links between the properties of items in an instrument, individuals responding to the items and the underlying constructs being measured, using a group of mathematical models. This justifies why we used the framework.

The research findings offer potential for actionable insight and have implications for educational practice. Estimating the psychometric properties of examination questions enhances the understanding of how the dimensionality, item difficulty, item discrimination, guessing parameters, students' ability estimates and reliability impact the validity and reliability of assessments. This knowledge informs teachers, researchers, test developers and evaluators in designing more effective and fair examinations, and ultimately, improve the quality of education and the evaluation process. From the findings of this study, the following recommendations are made:

1) The number of dimensions that account for the variations in student responses in an examination should always be determined. This will enhance an appropriate interpretation of scores and comparison of students' performances as well as making good decisions about the students and the educational system at large.
2) The quality of each item in a test should always be determined and should be directly proportional to the quality of the test as a whole.
3) IRT is highly recommended as a framework to determine these item parameters because of its numerous advantages over the CTT.

Acknowledgements

We are grateful to all our colleagues and the authors of whom studies were consulted in the course of this study.

Authors' Contributions

JJA – conceptualization, data curation, formal analysis, methodology, project administration, resources, validation, visualisation, writing original draft, writing review and editing. OCZ – data

curation, methodology, project administration, resources, validation, visualisation, writing original draft, writing review and editing. BCEO – conceptualisation, data curation, formal analysis, methodology, project administration, resources, validation, visualisation, writing original draft, writing review and editing. UUJ – data curation, methodology, project administration, resources, validation, visualisation, writing original draft, writing review and editing. AIN – data curation, methodology, project administration, resources, validation, visualisation, writing original draft, writing review and editing. FCO – data curation, methodology, project administration, resources, validation, visualisation, writing original draft, writing review and editing. CUE – data curation, methodology, project administration, resources, validation, visualisation, writing original draft, writing review and editing. SUN – data curation, methodology, project administration, resources, validation, visualisation, writing original draft, writing review and editing. TOU – data curation, methodology, project administration, resources, validation, visualisation, writing original draft, writing review and editing.

Notes

i. Published under a Creative Commons Attribution Licence.
ii. DATES: Received: 27 March 2022; Revised: 31 January 2025; Accepted: 21 June 2025; Published: 31 August 2025.

References

Abubakar MA & Usman HD 2021. The role of examination bodies in maintaining the credibility of high-stakes assessments in Nigeria. *Educational Measurement and Policy Review*, 18(3):200–215.

Adebayo OJ & Fakorede AM 2021. Analyzing the psychometric properties of Chemistry examination items: Implications for secondary education in Nigeria. *Journal of Educational Measurement and Evaluation*, 28(2):145–159.

Adonu IG 2014. Psychometric analysis of WAEC and NECO practical physics test using partial credit model. PhD thesis. Nsukka, Nigeria: University of Nigeria.

Akpan IE & Umoren DO 2018. The impact of instructional strategies on students' academic achievement in Chemistry. *Nigerian Journal of Science Education*, 22(3):52–67.

Ali A 2006. *Conducting research in education and social sciences*. Enugu, Nigeria: Tashiwa Networks.

Ani EN 2014. Application of item response theory in the development and validation of multiple-choice test in economics. PhD thesis. Nsukka, Nigeria: University of Nigeria.

Barton MA & Lord FM 1981. An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, 1981(1):i–8. https://doi.org/10.1002/j.2333-8504.1981.tb01255.x

Bereby-Meyer Y, Meyer J & Flascher OM 2002. Prospect theory analysis of guessing in multiple choice tests. *Journal of Behavioral Decision Making*, 15(4):313–327. https://doi.org/10.1002/bdm.417

Boughton KA & Yamamoto K 2007. A hybrid model for test speededness. In M von Davier & CH Carstensen (eds). *Multivariate and mixture distribution Rasch models: Extensions and applications*. New York, NY: Springer. https://doi.org/10.1007/978-0-387-49839-3_9

Chikezie I 2017. Assessment of unidimensionality of West African Senior School Certificate Examination in Chemistry with participation component analysis and item response theory model. *African Journal of Theory and Practice of Educational Assessment*, 5(6):46–57.

Collins SW, Chambers TG & Prather EE 2018. *An item response theory evaluation of the Light and Spectroscopy Concept Inventory national data set*. USA: National Science Foundation Collaboration of Astronomy Teaching Scholars (CATS). Available at https://arxiv.org/pdf/1709.05255.pdf. Accessed 18 December 2021.

De Ayala RJ 2009. *The theory and practice of item response theory*. London, England: The Guilford Press.

Delavar A & Zahrakar K 2013. *Assessing and measuring in psychology and educational sciences*. Tehran, Iran: Arasbaran.

DeMars C 2010. *Item Response Theory: Understanding statistics measurement*. New York, NY: Oxford University Press.

Donald A, Jacobs LC, Razavieh A & Sorensen CK 2009. *Introduction to research in education* (8th ed). Belmont, CA: Cengage Learning.

Ene C 2014. Development and calibration of a Basic Science achievement test using the two-parameter logistic model of item response theory. PhD thesis. Nsukka, Nigeria: University of Nigeria.

Ezeugo IC & Onwuka UN 2022. Application of the multidimensional 4-parameter logistic model in educational assessment: A case study of chemistry examinations. *International Journal of Educational Research and Development*, 35(1):97–112.

Fraser C & McDonald RP 2003. *NOHARM 3.0*.

Hambleton RK & Swaminathan H 1981. Book review: Lord, F. M. *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erbaum Associates, 1980. xii + 274 pp $29.95. *Journal of Educational Measurement*, 18(3):178–180.

Hingorjo MR & Jaleel F 2012. Analysis of one-best MCQs: The difficulty index, discrimination index and distractor efficiency. *Journal of the Pakistan Medical Association*, 62(2):142–147. Available at https://www.researchgate.net/publication/228111127. Accessed 18 January 2022.

Jang EE & 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

Kolen MJ & Brannen RL 2014. *Testing equating, scaling, and linking: Method and practices* (3rd ed). New York, NY: Springer. https://doi.org/10.1007/978-1-4939-0317-7

Kose IA & Demirtasli NC 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

Lee C 2019. What is item analysis? And other important exam design principles: How item analysis can increase teaching efficacy and assessment accuracy. Turnitin, 10 September [Blog]. Available at https://www.turnitin.com/blog/what-is-item-analysis-and-other-important-exam-design-principles. Accessed 14 February 2022.

Magis D 2013. A note on the item information function of the four-parameter logistic model. *Applied Psychological. Measurement*, 37(4):304–315. https://doi.org/10.1177/0146621613475471

Magis D & Raîche G 2012. Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software*, 48(8):1–31. https://doi.org/10.18637/jss.v048.i08

Maniaci MR & Rogge RD 2014. Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, 48:61–83. https://doi.org/10.1016/j.jrp.2013.09.008

Maydeu-Olivares A 2015. Evaluating the fit of IRT models. In SP Reise & DA Revicki (eds). *Handbook of item response theory modeling: Applications to typical performance assessment*. London, England: Routledge. https://doi.org/10.4324/9781315736013

Miller MD, Linn RL & Gronlund NE 2009. *Measurement and assessment in teaching* (10th ed). Upper Saddle River, NJ: Merrill, Prentice Hall.

Mroch AA, Bolt DM & Wollack JA 2005. *A new multi-class mixture Rasch model for test speededness*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada, April. Available at https://testing.wisc.edu/research%20papers/NCME%202005%20paper%20(Mroch,%20Bolt,%20&%20Wollack).pdf. Accessed 28 March 2021.

Nwosu PO & Adesina MO 2020. Students' attitudes towards chemistry and their performance in high-stakes examinations: A correlational study. *Journal of Science Education*, 30(4):311–326.

Obiekwe OO & Okoye CO 2019. Trends in students' performance in Chemistry at WASSCE: A five-year analysis. *African Journal of Educational Research*, 25(2):89–102.

Obinne ADE 2012. Using IRT in determining test items prone to guessing. *World Journal of Education*, 2(1):91–95. https://doi.org/10.5430/wje.v2n1p91

Ogunleye AO & Babajide VFT 2019. The significance of Chemistry in secondary school curriculum and its implications for tertiary education. *Nigerian Journal of Curriculum Studies*, 26(1):105–119.

Okwilagwe EA & Ogunrinde MA 2017. Assessment of unidimensionality and local independence of WAEC and NECO 2013 geography achievement tests. *African Journal of Theory and Practice of Educational Assessment*, 5:31–45.

Olatoye RA & Aderogba AA 2019. Laboratory facilities and students' performance in Chemistry: A case study of selected secondary schools in Nigeria. *Journal of Science Teaching and Learning*, 24(2):77–90.

Olonade PO, Metibemu MA & Adewale JG 2017. Unidimensional item response theory versus multidimensional item response theory: Evaluating the similarity of item calibration results in mathematics test items in Lagos State, Nigeria. *African Journal of Theory and Practice of Educational Assessment (AJTPEA)*, 5(6):73–86.

Rae A 2014. Trend analysis. In AC Michalos (ed). *Encyclopedia of quality of life and well-being research*. Dordrecht: Springer. https://doi.org/10.1007/978-94-007-0753-5_3062

Sinharay S & Lu Y 2008. A further look at the correlation between item parameters and item fit statistics. *Journal of Educational Measurement*, 45(1):1–15. https://doi.org/10.1111/j.1745-3984.2007.00049.x

Skidmore 2017. *What is academic assessment?*

Sorum MJ 1958. The assumption of normality in statistical work. PhD dissertation. Minneapolis, MN: University of Minnesota.

Stout W 2005. *DIMTEST (Version 2.0)* [Computer software]. Champaign, IL: The William Stout Institute for Measurement.

Tommy UE & Udo EM 2019. Examining item difficulty and student ability parameters of National Examinations Council's Biology examinations using the Rasch measurement Model in Nigeria. *British Journal of Education*, 7(8):66–80. Available at https://eajournals.org/wp-content/uploads/Examining-Item-Difficulty-and-Student-Ability-Parameters-of-National-Examinations-Councils-Biology-Examinations-using-the-Rasch-Measurement-Model-in-Nigeria.pdf. Accessed 18 December 2021.

Umobong E & Tommy UE 2017. Dimensionality of National Examinations Council's Biology Examinations: Assessing test quality in modern trend approach. *African Journal of Theory and Practice of Educational Assessment*, 5(2):14–30.

Van der Linden WJ 2007. A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72(3):287–308. https://doi.org/10.1007/s11336-006-1478-z

Yeh CC 2007. The effect of guessing on assessing dimensionality in multiple-choice tests: A Monte Carlo study with application. PhD dissertation. Pittsburgh, PA: University of Pittsburgh. Available at https://www.proquest.com/docview/304822013?pq-origsite=gscholar&fromopenview=true&sourcetype=Dissertations%20&%20Theses. Accessed 21 August 2025.

Zhang J 2006. Conditional covariance theory and detect for polytomous items. *Psychometrika*, 72(1):69–91. https://doi.org/10.1007/s11336-004-1257-7

Ziegler M & Hagemann D 2015. Testing the unidimensionality of items: Pitfalls and loopholes. *European Journal of Psychological Assessment*, 31(4):231–237. https://doi.org/10.1027/1015-5759/a000309

Zondo NP, Zewotir T & North DE 2021. The level of difficulty and discrimination power of the items of the National Senior Certificate Mathematics Examination. *South African Journal of Education*, 41(4):Art. #1935, 13 pages. https://doi.org/10.15700/saje.v41n4a1935