ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(1); 1-18

Differential Bundle Functioning of National Examinations Council Mathematics Test Items: An Exploratory Structural Equation Modelling Approach

Oluwaseyi Aina OPESEMOWO* Musa Adekunle AYANWALE**
Titilope Racheal OPESEMOWO*** Eyitayo Rufus Ifedayo AFOLABİ****

Abstract

A differential bundle function (DBF) is a situation in which examinees who are of the same ability but are from different groups are required to answer groups of items differently. DBF undermines the validity of the instrument if inadequately considered. The study examines the dimensionality of the 2017 NECO Mathematics items, determines the effect of DBF on 2017 Mathematics items concerning sex, and investigates the effect of DBF on 2017 Mathematics items concerning school ownership. This study explores Exploratory Structural Equation Modelling (ESEM), which permits the cross-loading of items that are not allowed in other models. The ex-post facto research design was adopted using secondary data, while six bundles were generated via the specification table. The population for the study comprised all 1,034,629 Senior School three students. A total of 815,104 students were selected using the simple random technique. The instrument for the study was 2017 NECO Mathematics paper three with a Cronbach's alpha of 0.82, and data were analysed using Mplus 7.4. The results revealed that the 2017 NECO Mathematics is multidimensional and items in the bundles possess construct validity as they functioned differentially to examinees' sex and school type. We recommend ESEM has a better approach to examining DBF on 2017 NECO Mathematics test items.

Keywords: Differential bundle functioning, exploratory structural equation modelling, National Examinations Council

Introduction

The dismal performance of examinees in the Senior School Certificate Examination (SSCE) could be a result of differential bundle performance that is spotted among examinees' group (male/female), and this (dismal performance) can lead to item bundle bias. To ensure that item bundles are fair to all intended groups, examination bodies should modify or delete bundles that may flag Differential Bundle Functioning (DBF) across examinees' groups. The instrument to measure the ability of groups of examinees becomes unfair when DBF occurs. There are four ways by which test fairness is categorised as submitted by standards for Educational and Psychological Testing (Boughton et al., 2000). Firstly, a fair test must be free from bias. Bias occurs when tests yield or promote scores that result in different meanings for members of different groups of examinees with the same competence level. Secondly, test fairness requires that examinees have received equal justice and treatment in the testing process. The achievement of fair treatment in a standardised test can be actualised when awarding scores to individuals and examinees groups by considering the items in the test and the testing context. Third of them is that test outcomes must be equitable to ensure test fairness, meaning examinees must have an equal opportunity to demonstrate proficiency in the measured construct. Examinees with the same

* Postdoctoral Research Fellow, University of Johannesburg, Faculty of Education, Johannesburg-South Africa, opesemowoo@gmail.com, ORCID ID: 0000-0003-0242-7027
** Senior Research Fellow, University of Johannesburg, Faculty of Education, Johannesburg-South Africa, ayanwalea@uj.ac.za, ORCID ID: 0000-0001-7640-9898
*** Research Assistant, Obafemi Awolowo University, Faculty of Education, Ile Ife-Nigeria, oluwatimilehint@gmail.com, ORCID ID: 0000-0002-0553-7355
**** Prof., Obafemi Awolowo University, Faculty of Education, Ile Ife-Nigeria, eriafolabi@gmail.com, ORCID ID: 0000-0002-0014-0711

To cite this article:
Opesemowo, O. A., Ayanwale, M. A., Opesemowo, T. R., & Afolabi, E. R. I. (2023). Differential bundle functioning of National Examinations Council mathematics test items: An exploratory structural equation modelling approach. *Journal of Measurement and Evaluation in Education and Psychology, 14*(1), 1-18. https://doi.org/10.21031/epod.1142713

Received: 9.7.2022
Accepted: 8.11.2022

capacity should receive the same score if there is no bias in the testing process. Lastly, to be fair to examinees, it is important to ensure they have had the opportunity to learn the content covered in the achievement domain during the preparation for the exam (Boughton et al., 2000). Summarily, the multifaceted nature of test fairness has made it practically impossible to have a generally accepted definition. When groups or bundles of items are unfair in measuring the same construct, such groups or bundles of items reflect DBF.

DBF occurs after controlling for the overall capacity of the construct measured by the test for examinees with the same ability but belonging to different groups who have a different probability of answering groups of items correctly. Item bias and impact may be responsible for DBF (Latifi et al., 2016). More so, when it involves two groups, such as examinees from rural/urban communities, male/female examinees, or public/private school students with the same ability, one would expect that examinees receive a similar score on a particular bundle of items. The comparability of test results across cultures has also been investigated using DBF (Ong et al., 2015). When one group persistently receives a lower score on an item bundle because of insufficient knowledge to answer items correctly in the bundle or something other than the knowledge of the subject matter being measured, DBF is said to have taken place.

Similarly, examinees of the same ability in this comparison group (male/female, urban/rural) are expected to answer the clustered items correctly and receive the same score for the correctly answered item bundle. But when the contrary occurs, bias is said to have been introduced against a particular group. Furthermore, for the bundle of items to measure mathematics ability correctly, individuals who have similar knowledge and expertise should have an equal chance of getting the answer correctly. When people with the same capacity in different groups, say male and female have a different probability of successfully answering an item, that item is said to function differently (differential performance) across groups (Ong et al., 2015). Differential performance may be assessed for an individual item called Differential Item Functioning (DIF). However, DIF occurs when an examinee with the same ability but belongs to different groups has a different probability of answering an item correctly.

In contrast, when it involves groups of items measuring the same construct and examinees with the same ability have a different chance of responding correctly to such bundles, DBF occurs. When the probability or chance to answer bundles of items rightly differs from examinees with the same ability level but belong to a separate group, DBF takes place (Min & He, 2020). The concept of DBF was built upon the DIF. In the DBF, items will be categorised into bundles and crisscross whether any item in the bundle demonstrates differential performance. To bundle items, various researchers (Furlow et al., 2009; Gierl et al., 2005; Oshima et al., 1998) have outlined diverse methods to group items into bundles.

Item Bundle Creation

Item bundle is a set of test items that are supposed to measure a universal secondary dimension e.g., items measuring the same construct. In addition, a bundle is a set of items measuring the same construct or the measurement of groups of items to test a particular learning domain (cognitive, affective, and psychomotor). An established principle guides item bundle creation. The DBF impacts ability estimation in no small measure, which is why bundling items suspected of DBF is crucial. If such DBF is not detected, there can be bias in both measurements and ability estimations. A bundle can be created using several organizing principles. These include a table of specifications, expert knowledge, and statistical detection. Based on the test's different content and cognitive dimensions, the table of specifications indicates a multidimensional structure in the data. Thus, items can be sampled from this specification table to determine whether different content areas have multiple dimensions. Expert knowledge is a confirmatory approach (Douglas et al., 1996). To use this method, content experts will be required to identify groups of items that are believed to measure the same construct in the test. With this method, the expert will examine each item and then determine if the items share a common theme or similar content to make bundles to test for DBF based on these themes or similarities in content. The exploratory approach to bundling items has also been proposed by (Douglas et al., 1996). This approach involves

using statistical procedures to identify distinct dimensions; however, various analytical methods are available for structuring items as a group, such as factor analysis, cluster analysis, and multidimensional scaling, to name just a few.

After reviewing the methods of creating bundles, the researchers adopted a table of specifications in this study. Likewise, various statistical methods such as the Simultaneous Item Bias Test (Shealy & Stout, 1993; Walker et al., 2011) and Multiple Indicator Multiple Causes (Finch, 2012; Lee et al., 2016; Montoya & Jeon, 2019; Mucherah et al., 2012) have been used to investigate DBF. The Exploratory Structural Equation Modeling (Asparouhov & Muthen, 2009) was adopted in this study. In Africa and Nigeria, various research studies have been conducted on DIF, but there has been a dearth of research on DBF. Consequently, DBF poses significant threats to item and bundle parameters that inform sex and other examinees' characteristics on NECO Mathematics test item performance. Such risks reflect noticeably on examinees' performance in Mathematics and may be responsible for the dismal performance commonly reported; therefore, the need arises to illuminate this threat of DBF using the Exploratory Structural Equation Modeling (ESEM) approach. It is imperative to state that this study employs the ESEM to determine if DBF exists among examinees' sex and school type in NECO Mathematics items. This examination is peculiar because all students who aspire to proceed to the higher institution of learning in Nigeria must pass the certification examination.

Appraisal of ESEM

Before the introduction of ESEM (Asparouhov & Muthén, 2009; Marsh et al., 2014; Morin & Maïano, 2011), the Exploratory Factor Analysis (Jennrich & Sampson, 1966), and Confirmatory Factor Analysis (Jöreskog, 1969) have been used to test factor, convergent, and divergent scores of numerous psychological instruments. Exploratory Factor Analysis (EFA) seeks to uncover the underlying structure of a relatively large set of variables. This was the first technique that was commonly used for factor analysis. The EFA is applied at an early stage of instrument development, and at this stage, there is the nonappearance of the structural specification of the instrument (Tsigilis et al., 2018). The researcher described this as a data-driven technique (Brown, 2015). The EFA has some limitations. One of these limitations is that it does not include technique effect adjustments. For instance, two items with similar wording can appear in an instrument. Therefore, we often need to include residual correlations to explain the covariance of these items with their latent constructs. Researchers often take into account the comparison between scores obtained from different groups of participants when analysing instrument scores, as well. It can be concluded, therefore, that the comparison scores have meaning only if the same number is interpreted the same way for all the groups in a particular study.

On the other hand, according to pre-established theory, a Confirmatory Factor Analysis (CFA) attempts to determine whether the number of factors and the loading of measured variables corresponds to what was expected based on the number of factors and loadings. The CFA is based on the assumption that items are loaded onto their respective factors, and those cross-loading items onto one or more latent factors are not permitted. An instrument structure in CFA can be determined by looking at the theoretical assumptions as well as the outcomes of previous EFAs or results that were produced. CFA has therefore been described as a methodological approach that is conceptually driven (Tsigilis et al., 2018). It is worth noting that the advantage of CFA is its ability to elucidate whether and to what extent the measurement model generalises across groups, as well as the relative consistency of the scores obtained. Several modifications have been introduced into the exploratory model to improve the model's fit by examining certain aspects of the model that are ill-fit to the data. The ESEM method was developed by Asparouhov and Muthén (2009) as an improvement to the previous approach. The ESEM allows the user to combine both the EFA and CFA in one model, providing a holistic framework that allows both to be used simultaneously. There is no doubt that ESEM is an improvement on EFA and CFA in that it combines both improvements into a single framework where factors will cross-load at some point. ESEM also has the advantage of allowing the simultaneous analysis of all cross-loadings in the form of a single cross-loading at a time, which can be calculated based on the modifications indices of the cases that have been analysed (Morin & Maïano, 2011) in a single step. More importantly, when compared with EFA and CFA, ESEM is much more accurate at fitting the data to the model when compared to

both. So far, several studies have been conducted on DBF using other models in other countries. Still, there is a paucity of research on DBF using ESEM in Nigeria to the best of our knowledge.

Purpose of the Study

Examinees continue to perform dismally in NECO mathematics, and major stakeholders in the education industry continue to pay attention to the issue. It has been attempted by several researchers to identify the factors responsible (such as shortage of qualified teachers (Ojimba, 2012), lack of equipment and instructional materials for effective teaching (Akale, 1997), poorly motivated teachers, and overcrowded classrooms (Asikhia, 2010), students' poor attitude towards mathematics (Akinsola, 1994), poor methods of teaching mathematics (Asikhia, 2010), poor learning environment (Black, 2001; Tata et al., 2014), students poor study habits and orientation (Aremu & Sokan, 2003; Umameh, 2011), school location (Adeyemo, 2005), lack of parental participation (Uwadie, 2012), gender of the teacher (Adeyegbe & Oke, 2002; Adeyemo, 2005), nature of the test items and examinees' characteristics (Ayanwale, 2019; Awopeju & Afolabi, 2016; Adeyemo & Opesemowo, 2020) for this performance of examinees in the external exam. There have also been several studies that suggest ways to improve student's math performance, such as mathematics can be taught in indigenous languages (Adegoke, 2011), improving instructional techniques (Abina, 2014), remunerating teachers well, and creating a conducive learning environment (Uwadie, 2012). Even though researchers have provided several interventions to improve performance, this dismal trend still persists. As a contributing factor, DBF was investigated in this study, which is different from what has been studied by others. There is a need to address bias since tests are used as gatekeepers for educational opportunities, and test items should be fair to all students. A test is relevant only if it produces valid outcomes for different subpopulations with the same measures. In addition, the importance of ensuring fairness and equity among examinees cannot be overstated. It is important to provide equal opportunities for all examinees to display their knowledge and perform well according to their demographic profiles (Ayanwale, 2022). When developing their test items, does NECO take this situation into consideration? To the best of the researcher's knowledge, this remains a mystery. When a test contains DBF elements, the student's performance will be adversely affected. Therefore, it is essential to examine the DBF of this public examining body from various demographic perspectives. Specifically, the purpose of this study was to ascertain the dimensionality of the 2017 NECO mathematics items, as well as determine the impact of DBF on 2017 NECO mathematics items based on school ownership and gender.

Research Questions

Research questions addressed in this study include:

1. What is the dimensionality of the 2017 NECO Mathematics items?
2. What is the effect of DBF on 2017 NECO Mathematics items concerning sex?
3. Is there any influence of DBF on 2017 NECO Mathematics items concerning school ownership?

Method

Design

This is quantitative research using ex-post facto design. As a design, ex post facto is known as "after-the-fact" research and examines how an independent variable (groups with certain qualities that already exist before a study) influences a dependent variable. As a result, a researcher cannot modify or manipulate actions or behaviours that have already taken place or specific traits and characteristics that a participant has (Creswell, 2003). Data were drawn from candidates' responses who wrote the 2017 NECO Mathematics paper three examinations. Mathematics paper 3 consists of multiple-choice items. The population for the study comprised all of the 1,034,629 Senior School three (SS 3) students who

registered and took the examination. The population figure was made available in the data provided by NECO. The NECO Mathematics examination is a national examination usually administered annually and taken by all candidates in the 36 states, including the Federal Capital Territory (FCT), Nigeria.

Participants

A survey system sample calculator was used to determine the sample size. The sample size was set at a 95% confidence level and 0.05 confidence interval. The study sample consisted of 815,104 SS 3 students, 393,695 (48.3%) males, and 421,409 (51.7%) females were selected using a simple random technique. Also, 497 schools (i.e., 318 (63.98%) private and 179 (36.02%) public schools) across the six geopolitical zones in Nigeria that enrolled students for the NECO Mathematics examination were selected using purposive sampling techniques. Data were retrieved from the Optical Mark Recorder (OMR) sheets, obtained from the NECO head office, Minna, Niger State, Nigeria.

Instrument

The 2017 June/July NECO SSCE Mathematics paper three examination was the instrument. It was a dichotomous (i.e., correct response scored one while wrong response scored zero) multiple-choice examination comprising 60 items with a key and four distracters making five alternative responses, and the items were based on the Senior Secondary School (SSS) Mathematics curriculum. Examinees had to provide information about themselves, such as sex, location, name, examination number, school name, serial number, and subject code. The response options for the instrument range from letters A-E. After the third year of SSS, the SSCE is usually administered. In much the same way, the NECO exam serves as an assessment mechanism that ascertains the extent to which a student has acquired essential skills and competencies. A specifications table (Table 1) showed how items were distributed across the behavioural objectives and contents. The instrument has a Cronbach alpha of 0.89 reliability.

The table of specifications (Table 1) demonstrated that knowledge possesses seven items representing 11.6%, comprehension had six items with 10%, the analysis had 22 items representing 36.7%, the application had 16 items cum 26.7%, synthesis had five items representing 8.3%, and evaluation had four items representing 6.7%. Also, it may deduce that analysis revealed the highest number of items while evaluation had the least items. Similarly, number and numeration showed 11 items representing 18.3%, algebra had 18 items accounting for 30%, mensuration showed 6 items representing 10%, geometry had 9 items representing 15%, statistics and probability had 10 items showing 16.7% and introduction to calculus had 6 items which accounted for 10%. Additionally, it was shown that algebra had the highest number of items, while mensuration and introduction to calculus possessed the least number of items.

Data Analysis

Data obtained was analysed using Mplus software version 7.4 (Muthén & Muthén, 2012) and estimated with the robust maximum likelihood estimator (MLR), which provides standard errors and tests of model fit that are robust to the non-normality of the data. Also, examinees' responses were subjected to the Stout test of essential unidimensionality (Stout, 1987), a nonparametric analysis using DIMTEST package.

Table 1

Table of Specification of the 2017 NECO Mathematics Items

Content	**Cognitive Skills**						**Total**
	Know. (11.6%)	Comp. (10%)	Ana. (36.7%)	App. (26.7%)	Syn. (8.3%)	Eva. (6.7%)	
N/N (18.3%)	0	2(ITs 20 & 33)	1(IT 9)	2(ITs 5 & 10)	4(ITs 1,3,4&11)	2(ITs 2&6)	11
ALG. (30%)	1(IT 24)	3(ITs 18,31&48)	6(ITs 8,15,19,23,30&45)	8(ITs 7,12,13,16,21,22,25&32)	0	0	18
MEN. (10%)	0	0	4(ITs 37,38,39, &42)	2(ITs 17&43)	0	0	6
GEO. (15%)	0	0	8(ITs 34,35,36,40,41,44,46&51)	1(IT 48)	0	0	9
STAT/PROB. (16.7%)	6(Its 26,27,50,52&53)	1(ITs 49)	3(ITs 54,55&56)	0	0	0	10
INTRO. TO CAL. (10%)	0	0	0	3(ITs 14,28&29)	1(IT 59)	2(ITs 58&60)	6
TOTAL	7	6	22	16	5	4	60

Note. IT = Item; ITs = Items; N/N = Number and Numeration; ALG = Algebra; MEA= Mensuration; GEO = Geometry; STAT/PROB = Statistics and Probability; INTO. TO CAL. = Introduction to Calculus; Know.: Knowledge; Comp. = Comprehension; Ana. = Analysis; App. = Application; Syn. = Synthesis; Eva. = Evaluation

Model Fit Statistics

Several model fit statistics have been used by researchers to assess structural equation models, but in this study, the researchers considered chi-square statistic, Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and Root Mean Square Error of Approximation (RMSEA) provided in Mplus (Muthén & Muthén, 2012). The model chi-square ($\chi 2$) statistic is the original fit index for the structural model (Wang & Wang, 2012), which assesses the overall fit and discrepancy between the sample and fitted covariance matrices. When $\chi 2$ is significant, the null hypothesis will be rejected such that the model fits the population and vice-visa. The degree of freedom (df) is the discrepancy between the number of pieces of available information less the number of estimated model parameters. The chi-square statistic is expressed by

$$\chi^2 = f_{ML}\,(N-1) \tag{1}$$

where fML = $(S, \Sigma^{\wedge})$ is the model-specific minimum fit function value, and N is the sample size. The cut-off for a good fit is p-value > 0.05

Comparative Fit Index (CFI): The CFI belongs to the group of incremental fit indices that compare the fit of the hypothesised model to the fit of a baseline model, which is the independent model. It demonstrates how much the hypothesised model fits better than the more parsimonious independence model. The CFI is a measure based on the noncentrality parameter d = ($\chi 2$ – df) where df represents the degree of freedom of the model. The formula for the CFI is

$$\text{CFI} = \frac{d_{null} - d_{specified}}{d_{null}} \tag{2}$$

where d_{null} and $d_{specified}$ are the noncentrality parameters for the null model and the specified model. The rule of thumb cutoff for the CFI is 0.90 but increased from 0.90 to 0.95 by Hu and Bletter (1999).

Tucker-Lewis Index (TLI): Another way to compare the lack of fit of a specified model to the lack of fit of the null model is the use of TLI. The TLI is also an incremental fit index, which does not guarantee a value from 0 to 1 and compares the fit of the target model to the fit of the independence model. The TLI is also known as the Non-Normed Fit Index (NNFI) and it is expressed as:

$$\text{TLI} = \frac{\left(\frac{\chi^2_{null}}{df_{null}} - \frac{\chi^2_{specified}}{df_{specified}}\right)}{\left(\frac{\chi^2_{null}}{df_{null}} - 1\right)} \tag{3}$$

where χ^2_{null} / df_{null} and $\chi^2_{specified}$ / $df_{specified}$ are ratios of χ^2 statistics to the degrees of freedom of the null model and the specified model, respectively. TLI has punishment for model complexity because the freer parameters, the smaller the $df_{specified}$, thus the larger $\chi^2_{specified}$ / $df_{specified,}$ leading to a smaller TLI model (Wang & Wang, 2012). A value greater than 0.95 has been regarded as the rule of thumb or cut-off criteria.

Root Mean Square Error of Approximation (RMSEA): The coefficient of the RMSEA is used to measure the approximate model fit. It is based on the non-centrality parameter:

$$\text{RMSEA} = \sqrt{\frac{(\chi_S^2 - df_S)}{df_S}} = \sqrt{\frac{(\chi_S^2/df_S) - 1}{N}} \quad (4)$$

where $(\chi_S^2 - df_S)$/N is the rescaled noncentrality parameter to adjust for sample size. It is understood that RMSEA values range from 0 to 0.10, where 0 indicates perfect fit, < 0.05 indicates close fit, 0.05-0.08 implies fair fit, 0.08-0.10 implies mediocre fit, and > 0.10 indicates poor fit (Browne & Cudeck, 1993; MacCallum et al., 1996; Byrne, 1998). A good model fit is defined as RMSEA ≤ 0.06 by Hu and Bentler (1999).

In view to assessing the dimensionality of the instrument, examinees' responses were subjected to the Stout test of essential unidimensionality (Stout, 1987), a nonparametric analysis using the Dimensionality Test (DIMTEST) package.

Results

Dimensionality Assessment

In psychological data involving subscales, one of the critical aspects that should be taken into account is the dimension of the data. A tenable assumption of unidimensionality needs to be made in any Item Response Theory (IRT) research context. There might be some degree of multidimensionality implied by an instrument with subscales. Tate (2002) states that when an instrument includes a subscale, there must be two aspects of validity that must be considered from a validity standpoint, namely the validity of the instrument's internal structure and the validity of the subscale's discriminant validity. Considering the assumption of unidimensionality, the first argument can be made. A dimensionality analysis should be performed before a DBF is assessed to ensure that the data are reasonably unidimensional (McCarty et al., 2007). In addition, dimensionality assessments can be useful in tests to determine whether or not the unidimensionality assumption used in the Unidimensional Item Response Theory (UIRT) has been strongly violated and may be used to measure whether or not this assumption has been at odds with the experimental results. Nevertheless, if evidence suggests that the unidimensionality of an item response theory is violated, then alternative methods can be used to find scores, such as those based on the Multidimensional Item Response Theory (MIRT). It is also possible to make predictions using the dimensionality assessments of different bundles of data as well as determine how those results can be compared with each other from different bundles of data.

Research Question One: *What is the dimensionality of the 2017 NECO Mathematics items?*

The dimensionality could either be unidimensional or multidimensional. However, the Stout test of essential unidimensionality is obtainable by dividing the items into two different groups. The first group of items consists of the Assessment Subtest (AT), which is designed in a way that is homogeneous with the rest of the group while also being dimensionally different from the remainder of the items in the group. There is a second group of items known as the Partitioning Subtest (PT). These are items that are not included in the AT. The grouping of items into two can be done by adopting either exploratory or confirmatory analysis but in this study, the exploratory analysis in DIMTEST was implemented.

Table 2

The Dimensionality of 2017 NECO Mathematics Items

TL	TGbar	T	p-value
33.04	13.68	16.89	0.00

The result of the test of Stout's essential unidimensionality (Table 2) was used to investigate the assumption of unidimensionality of the instrument that might form a secondary dimension. Two subtests, AT and PT, were divided into the test. A dominant trait is chosen as the item that measures the dominant trait and the AT in the most effective way. It seems that these items measure best when measured in a direction distinct from the direction of the PT items. It was decided to use the HCA/CCPROX clustering procedure to select the AT and the DETECT statistics in DIMTEST to perform the analysis. These items' cluster was tested to ascertain if it was dimensionally distinct from the secondary dimension of the test. A random sample of 30% of the examinees' responses was used to select the AT (items clustered in AT are 1, 9, 10, 11,12, 14, 16, 19, 21, 24, 28, 30, 31, 32, 33, 34, 35, 38, 39, 41, 43, 45, 46, 48, 49, and 50), and the remaining 70% of the examinees' responses were used as PT. The null and alternative hypotheses were tested using DIMTEST as proposed by Stout (1987).

H_0: $AT \cup PT$ satisfies essential a unidimensionality ($d = 1$)

H_i: $AT \cup PT$ fails to satisfy $d = 1$

Both the AT and PT assess a dimension that is dominant in the null hypothesis, while the items in the AT partition are best described by a dimension unique from the items in the PT partition. There was a violation of the essential unidimensionality assumption in the mathematics test items ($T = 16.87$, $p = 0.00$), resulting in the null hypothesis being rejected (Table 2). In addition, Table 2 presented the conclusion that the variance in the responses to the questions observed in the tests was attributable to multiple dimensions rather than one, which was the case previously. As a result of the implication of the above, there was a violation of the unidimensionality assumption involving the 2017 Mathematics items. This means that the 2017 NECO Mathematics item has a multifaceted aspect that must be considered. A further indication of the multidimensional nature of the 2017 NECO Mathematics items was provided by the T value, which was found to be statistically significant. It has been suggested by Furlow et al. (2009) that the use of UIRT models with multidimensional test data can violate or contradict the notion that all test items are equally dimensional and that there may be a potential hazard in estimating item and bundle parameters.

To address the research objectives, parameters were estimated with ESEM in Mplus 7.4 (Muthén & Muthén, 2012). Although cross-loading items are more visible and practicable with EFA, it is crystal clear that better techniques and approaches are more evident with CFA than with EFA. ESEM integrates the advantages of EFA and CFA into its technique. Thus, researchers such as (Ayanwale, 2022; Sass, 2011; Schmitt, 2011) argued that ESEM was a better and more efficient method to adjust for cross-factor loading instead of latent variables analysis, which assesses a measurement model of constructs through CFA. The model fit was established using chi-square ($\chi 2$), the CFI, TLI, and RMSEA.

Research Question Two: *Is there any statistically significant effect of DBF on 2017 NECO Mathematics with respect to sex?*

To answer this research question, the content analysis (Table 1) was developed as items were set into different bundles by implementing the confirmatory approach, and ESEM was adopted in analysing the data. The result is presented in Table 3.

Table 3

Differential Bundle Functioning of 2017 NECO Mathematics Items with Respect to Sex

Bundle	Estimate	S.E.	Est./S.E.	P-Value
1	0.113	0.006	18.833	0.000
2	0.121	0.007	17.286	0.000
3	0.093	0.008	11.625	0.000
4	0.102	0.007	14.571	0.000
5	0.087	0.007	12.429	0.000
6	0.088	0.008	11.000	0.000

Note: S.E. = Standard Error; Est. = Estimate

Table 4

Summary of Model Fit of ESEM

χ^2	df	RMSEA	CFI	TLI	p
169573.408	1749	0.043	0.964	0.958	0.0000

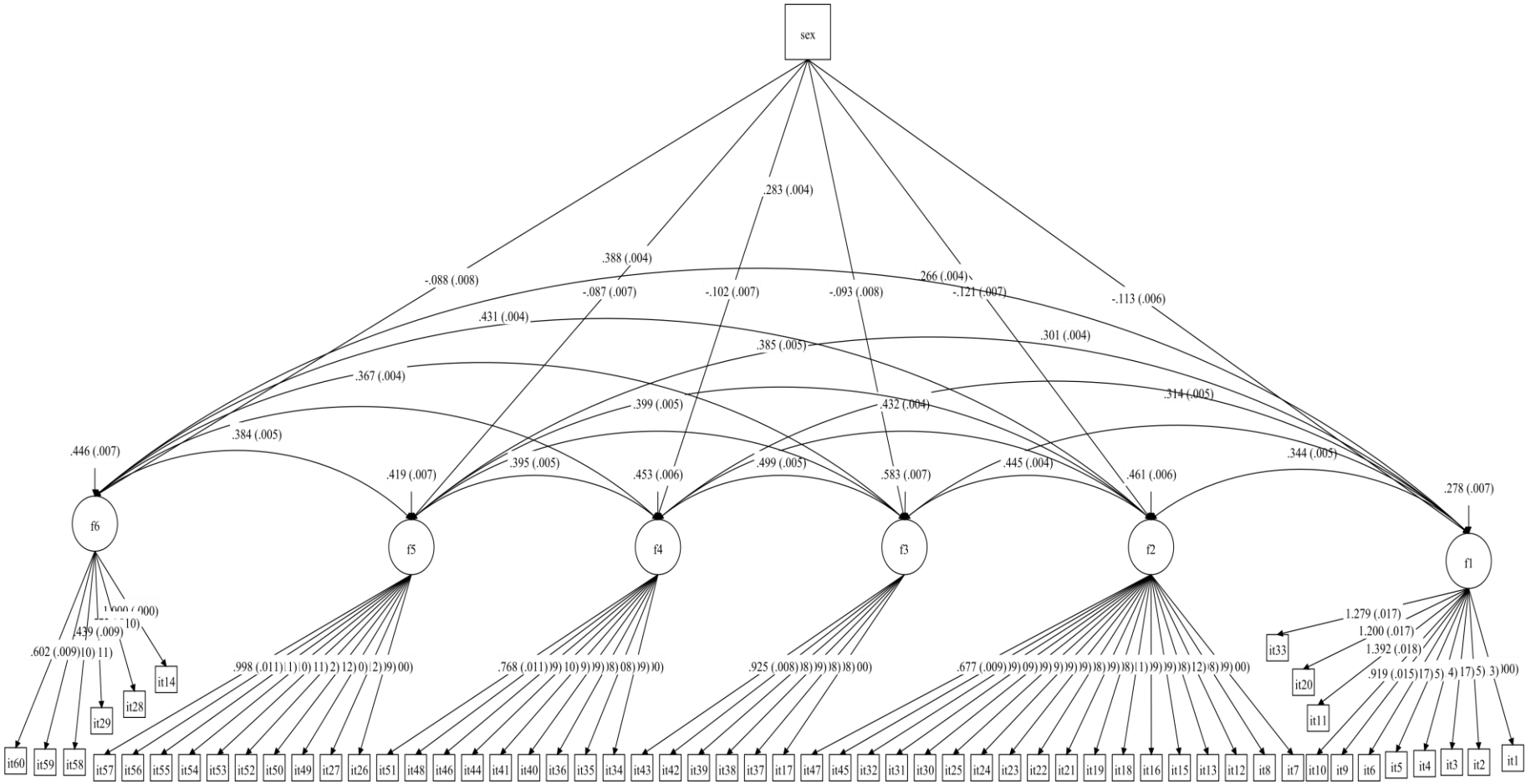
Note: RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index

The result (Table 3) showed that the items in the bundles possess construct validity as its fundamental factor functioned differentially with respect to examinees' sex. The result further displayed that all six bundles had a statistically significant effect on the examinees' sex with the underlying factor. the underlying factor of bundle 1, bundle 2, bundle 3, bundle 4, bundle 5, and bundle 6 are Z = 18.833, p = 0.000; Z = 17.286, p = 0.000; Z = 11.625, p = 0.000; Z = 14.571, p = 0.000; Z = 12.429, p = 0.000; Z = 11.000, p = 0.000 respectively. Table 4 further buttresses that the item bundles have construct validity as its fundamental factor functioned differentially with respect to examinees' sex. Also, the ESEM model was viable $\chi 2$ = 169573.408, df = 1749, p = 0.0000; RMSEA = 0.043 (90% C.I. = 0.043-0.043, probability of RMSEA $\leq$ 0.05 = 0.000), CFI = 0.964. TLI = 0.958. It also demonstrated the cross-loadings between all the items in the six bundles. However, the 2017 NECO Mathematics items at different bundles functioned differentially with respect to examinees' sex.

Figure 1 showed the model structure with estimated parameters of all the items and bundles in ESEM with respect to sex. It also demonstrated the cross-loadings between all the items in the six bundles. However, the 2017 NECO Mathematics items at various bundles functioned differentially with respect to examinees' sex.

Figure 1

Model Structure of Estimated Parameters with Respect to Sex



Research Question Three: *Is there any influence of DBF on 2017 NECO Mathematics items with respect to school type?*

To provide a valid answer to this research question, the content analysis (Table 1) was developed as items were grouped into bundles using the confirmatory approach based on items measuring the same construct, and ESEM was deployed in analysing the data. The results were presented in Table 5.

Table 5

Differential Bundle Functioning of 2017 NECO Mathematics Items with Respect to School Type

Variable	Estimate	S.E.	Est./S.E.	P-Value
Bundle 1	0.012	0.001	12.000	0.000
Bundle 2	0.009	0.001	9.000	0.000
Bundle 3	0.004	0.001	4.000	0.000
Bundle 4	0.006	0.001	6.000	0.000
Bundle 5	0.005	0.001	5.000	0.000
Bundle 6	0.005	0.001	5.000	0.000

Table 6

Summary of Model Fit Using ESEM

χ^2	Df	RMSEA	CFI	TLI	p
1235407.496	1830	0.043	0.964	0.958	0.0000

The result (Table 5) indicated that the item bundles pose construct validity as its fundamental factor functioned differentially with respect to the examinees' school type (public and private schools). The result also displayed that the bundles from bundles 1 to 6 had a statistically significant effect on the examinees' school type with the underlying factor. the underlying factor of bundle 1, bundle 2, bundle 3, bundle 4, bundle 5 and bundle 6 are Z = 12.000, p = 0.000; Z = 9.000, p = 0.000; Z = 4.000, p = 0.000; Z = 6.000, p = 0.000; Z = 5.000, p = 0.000 and Z = 5.000, p = 0.000 respectively. This implies (Table 6) that the item bundles have construct validity as its fundamental factor functioned differentially with respect to the examinees' school type. Also, the ESEM model (Table 6) was viable with $\chi 2$ = 1235407.496, df = 1830, p = 0.0000; RMSEA = 0.043 (90% C.I. = 0.043-0.043, probability of RMSEA $\leq$ 0.05 = 1.000), CFI = 0.964. TLI = 0.958. The data also had a good model fit as the CFI and TLI > 0.9. The differential performance noticed in the different bundles with respect to examinees' school type (public/private schools) may be attributed to the deficiency of an appropriate model for the test items, psychometric properties of the items not established, and lack of experience in the part of the item developer e.t.c.

Figure 2

Model Structure of Estimated Parameters with Respect to School Type

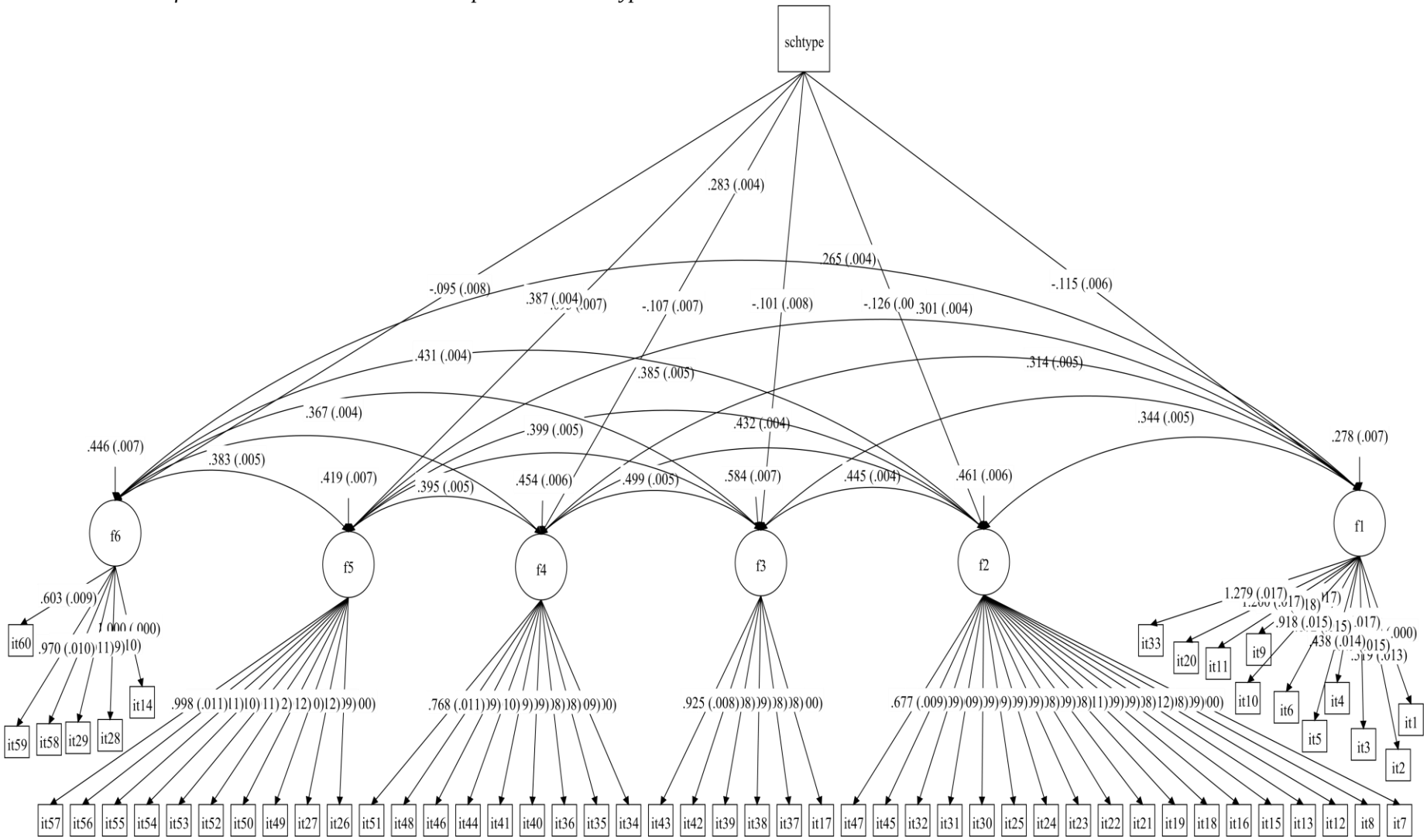


Figure 2 demonstrated the cross-loading of item bundles, and the bundles range from one to six. The cross-loading was achieved with the use of the ESEM. The differential performance noticed in the different bundles with respect to examinees' school type (public/private schools) may be attributed to the deficiency of an appropriate model for the test items, psychometric properties of the items not established, and lack of experience in the part of the item developer to mention but a few.

Discussion

This study aimed to ascertain the dimensionality of the 2017 NECO Mathematics items, determine the statistically significant effect of DBF of 2017 NECO Mathematics test items on examinees' sex, and finally, investigate the influence of DBF on 2017 NECO Mathematics items. However, the penultimate objective of the study was the use of the ESEM approach to determine the DBF of the 2017 NECO Mathematics test items. Furthermore, items were organised into various bundles using the confirmatory approach as postulated by Douglas et al. (1996). The 2017 NECO Mathematics test items had six bundles (Table 1) and each bundle was tested against the demographical variables (such as sex and school type) of the examinees. Based on the preliminary analysis of the study, it was revealed that the data is multidimensional and not unidimensional. Before DBF analysis, the dimensionality analysis must be established such that the data was reasonably unidimensional (Furlow et al., 2009). The application of UIRT models with multidimensionality data contradicts or violates the assumption of unidimensionality which invariably poses a statistically significant threat to bundle or item parameter estimates of examinees.

The dimensionality assessment of this study revealed the multidimensionality of the NECO Mathematics test items which was evident that more than one construct was measured. Similarly, suggestions have been made that tests like NECO comprise multiple-choice items, the different item types measure somewhat diverse traits, and therefore violate the IRT assumption of unidimensionality (Wainer & Thissen, 1993; Wainer et al., 1994). Also, it showed that more than one ability distribution is exhibited for an individual when the unidimensionality assumption of IRT has been compromised. When the assumption of unidimensionality has not been fulfilled, multidimensionality becomes the next alternative.

The second research question showed that there was a statistically significant effect of DBF on 2017 NECO Mathematics test items on the sex of examinees. This was in line with the study conducted by Boughton et al., (2000). They (Boughton et al., 2000) applied SIBTEST in understanding the differential performance of DBF on Mathematics and science achievement tests. They further revealed that male students consistently outperformed their female counterparts in Mathematics and Science. In addition, the result suggested that the model fit met the criteria postulated by Hu and Bentler (1999) that the CFI and TLI should be 0.90. ESEM uses either supplementary with CFA and it is an emerging technique used by researchers. Many studies (Marsh et al., 2020; Marsh et al., 2010; Marsh et al., 2009; Perry et al., 2015) conducted on ESEM have shown that ESEM is effective in the validation of a multidimensional measure like the 2017 NECO Mathematics test items. It was revealed in this study that the ESEM is a technique that is an appropriate substitute for CFA using Mathematics test items

The final research question also demonstrated a statistically significant influence of DBF on 2017 NECO Mathematics items with respect to the school type (public/private school) of the examinees. Walker et al. (2011) pointed out that the ability estimation bias can only be attributed to the DBF when a large number of items are showing whether focal groups of examinees perform differentially or not in a small way against that group of examinees or when a small number of items are showing differential performance against focal examinees in a large way. The existence of DBF in any standardised examination like NECO (which conducts a public examination that is used to adjudge whether a candidate is offered or denied admission into institutions of higher learning for Nigerian students) should be a cause of concern to the stakeholders in education. The essence is that the test scores obtained from the such national examination will be used to draw inferences about examinees' performance which will

invariably lead to overestimation or underestimation of examinees' ability thereby, leading to an erroneous judgment of the examinees' ability.

Based on the result, bundles feature items that share a common reading ability but may not share all cognitive tasks required for a correct response. The bundle of items associated with a mathematics test may be more difficult for an examinee who understands the term or does not understand the question. Also, the study finds that dependence within such bundles affects the distribution of items' responses in a predictable and testable manner. Although some small groups of items that share the same material exhibit excessive dependence, exam responses cannot be described as unidimensional. As a consequence, conventional IRT models can overestimate the standard error of measurement for exams with bundled items. Psychometric measurements for the six bundles are as follows: 0.51, 0.67, 0.68, 0.68, 0.65, and 0.44 respectively from bundles one to six. Some possible causes of those bundles of DBF in different groups of sex or school type might be the use of language structure of the items in the different bundle or when some items in the particular bundle focus on a certain area of interest like items relating to the sport. It is expected that male examinees might outperform their female counterparts in such items.

Conclusively, to ensure test fairness to all examinees, examination bodies like NECO should not only conduct DIF (DIF is not adequately proficient in detecting bias) but rather painstakingly apply ESEM which this study has shown to be effective in detecting DBF. Whenever DBF is detected, examination bodies are required to expunge or modify the item/item bundle (DBF) which can pose threat to the validity of an instrument which is the key focus of psychometricians.

Limitation to the study

The study was restricted to only NECO 2017 Mathematics test items, while further studies could be conducted on other subjects administered by NECO. A similar study should be conducted on various subjects of other public examining bodies such as West African Examination Council (WAEC), Joint Admission Matriculation Board (JAMB), and the National Business and Technical Examinations Board (NABTEB).

Declarations

Author Contribution: Oluwaseyi Aina Opesemowo: Conceptualization, methodology, analysis, writing & editing, visualization. Musa Adekunle Ayanwale: Methodology, analysis, writing & editing, visualization. Titilope Racheal Opesemowo: Conceptualization, methodology, writing & editing, visualization. Eyitayo Rufus Ifedayo Afolabi: Methodology, editing, and supervision.

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: Throughout this study, the researchers adhere to ethical principles. For the purpose of this study, secondary data were analyzed, which involved accessing pre-existing data that had been collected for research and anonymized. The study involved no direct participants, and no personal information was disclosed. This secondary data was used in accordance with the researchers' research plan and in a confidential and private manner. All relevant regulations, including institutional and national guidelines for data sharing and research ethics, were complied with by the researchers, who acknowledge and respect the efforts of those who collected and curated this dataset. Additionally, the researchers recognized that the data used represent individuals who might have contributed to their study, and they are committed to ensuring that their research will contribute to the advancement of psychometric knowledge. It is evident from the research conducted in this study that the researchers maintain the principles of respect for persons, beneficence, and justice and are committed to conducting their research in an ethical and responsible manner.

References

Abina, D. B. (2014). *Influence of teacher characteristics, availability and utilization of instructional materials on students' performance in mathematics* [Unpublished doctoral dissertation]. University of Ibadan.

Adegoke, B. A. (2011). Effect of direct teacher influence on dependent-prone students' learning outcomes in secondary school mathematics. *Electronic Journal of Research in Educational Psychology, 9*, 283-308.

Adeyegbe, S. O., & Oke, M. G. (2002). Science, technology and mathematics (STM) for sustainable development: The role of public examining bodies. *Proceedings of STAN annual conference* (pp. 144-147). Science Teachers Association of Nigeria.

Adeyemo, D. A. (2005). *Parental involvement interest in schooling and school environment as predictors of academic self-efficacy among senior secondary school students in Oyo State* [Unpublished doctoral dissertation]. University of Ibadan.

Adeyemo, E. O., & Opesemowo, O. A. (2020). Differential test let functioning (DTLF) in senior school certificate mathematics examination using multilevel measurement modelling. *Sumerianz Journal of Education, Linguistics and Literature, 3*(11), 249-253. https://doi.org/10.47752/sjell.311.249.253

Akale, M. A. G. (1997). The relationship between attitude and achievement among mathematics students in senior secondary school. *Journal of Science and Movement Education, 2*, 77-85.

Akinsola, M. K. (1994). *Comparative effects of mastery learning and enhanced mastery learning strategies on students' achievement and self-concept mathematics* [Unpublished doctoral dissertation]. University of Ibadan.

Aremu, O. A., & Sokan, B. O. (2003). *A multi-causal evaluation of academic performance of Nigerian learners: Issues and implication for national development*. Department of Guidance and counseling, University of Ibadan, Ibadan.

Asikhia, O. A. (2010). Students and teachers' perception of the causes of poor academic performance in Ogun state secondary schools: Implications for counseling for national development. *European Journal of Social sciences, 13*(2), 28-36.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 397-438. https://doi.org/10.1080/10705510903008204

Awopeju, O. A., & Afolabi, E. R. I. (2016). Comparative analysis of classical test theory and item response theory based item parameter estimates of senior school certificate mathematics examination. *European Scientific Journal, 12*, 263-284.

Ayanwale, M. A. (2019). *Efficacy of item response theory in the validation and score ranking of dichotomous and polytomous response mathematics achievement tests in Osun State, Nigeria* [Unpublished doctoral dissertation]. University of Ibadan.

Ayanwale, M. A. (2022). Performance of exploratory structural equation model (ESEM) in detecting differential item functioning. *EUREKA: Social and Humanities, 1*, 58–73. http://doi.org/10.21303/2504-5571.2022.002254

Black, S. (2001). Building blocks: How schools are designed and constructed affects how students learn. *American School Board Journal, 188*(10), 44-47.

Boughton, K. A., Gierl, M. J., & Khaliq, S. N. (2000). *Differential bundle functioning on mathematics and science achievement tests: A small step toward understanding differential performance*. Annual meeting of the Canadian Society for Studies in Education (CSSE), Edmonton, Alberta, Canada.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford publications.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Sage.

Byrne, B. M. (1998). *Structural equation modelling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Lawrence Erlbaum Associates.

Creswell, J. W. (2003). *Research design: Qualitative, quantitative and mixed methods approaches* (2nd ed.). Sage.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33*(4), 465-484. https://doi.org/10.1111/j.1745-3984.1996.tb00502.x

Finch, W. H. (2012). The MIMIC model as a tool for differential bundle functioning detection. *Applied Psychological Measurement, 36*(1), 40-59. https://doi.org/10.1177/0146621611432863

Furlow, C. F., Raiford, R. T., & Gagné, P. (2009). The impact of multidimensionality on the detection of differential bundle functioning using simultaneous item bias test. *Applied Psychological Measurement, 33*(6), 441-464. https://doi.org/10.1177/0146621609331959

Gierl, M. J., Tan, X., & Wang, C. (2005). *Identifying content and cognitive dimensions on the SAT®* (Report No. 2005-11). College Board.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Jennrich, R. I., & Sampson, P. F. (1966). Rotation for simple loadings. *Psychometrika, 31*(3), 313-323. https://doi.org/10.1007/BF02289465

Jöreskog, K. (1969). A general approach to confirmatory factor analysis. *Psychometrika, 34*, 183-202.

Latifi, S., Bulut, O., Gierl, M., Christie, T., & Jeeva, S. (2016). Differential performance on national exams: Evaluating item and bundle functioning methods using English, Mathematics, and Science Assessments. *SAGE Open, 6*(2), 1-14. https://doi.org/10.1177/2158244016653791

Lee, S., Bulut, O., & Suh, Y. (2016). Multidimensional extension of multiple indicators multiple causes models to detect DIF. *Educational and Psychological Measurement, 77*(4), 545-569. https://doi.org/10.1177/0013164416651116

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130-149. https://doi.org/10.1037/1082-989X.1.2.130

Marsh, H. W., Guo, J., Dicke, T., Parker, P. D., & Craven, R. G. (2020). Confirmatory factor analysis (CFA), exploratory structural equation modeling (ESEM), and set-ESEM: Optimal balance between goodness of fit and parsimony. *Multivariate Behavioural Research, 55*(1), 102-119. https://doi.org/10.1080/00273171.2019.1602503

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychological Assessment, 22*(3), 471-491. https://doi.org/10.1037/a0019227

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology, 10*(1), 85-110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 439-476. https://doi.org/10.1080/10705510903008220

McCarty, F. A., Oshima, T. C., & Raju, N. S. (2007). Identifying possible sources of differential functioning using differential bundle functioning with polytomously scored data. *Applied Measurement in Education, 20*(2), 205-225. https://doi.org/10.1080/08957340701301660

Min, S., & He, L. (2020). Test fairness: Examining differential functioning of the reading comprehension section of the GSEEE in China. *Studies in Educational Evaluation, 64*, 100811. https://doi.org/10.1016/j.stueduc.2019.100811

Montoya, A. K., & Jeon, M. (2019). MIMIC models for uniform and nonuniform DIF as moderated mediation models. *Applied Psychological Measurement, 44*(2), 118-136. https://doi.org/10.1177/0146621619835496

Morin, A. J. S., & Maïano, C. (2011). Cross-validation of the short form of the physical self-inventory (PSI-S) using exploratory structural equation modeling (ESEM). *Psychology of Sport and Exercise, 12*(5), 540-554. https://doi.org/10.1016/j.psychsport.2011.04.003

Mucherah, W., Finch, W. H., & Keaikitse, S. (2012). Differential bundle functioning analysis of the self-description questionnaire self-concept scale for Kenyan female and male students using the MIMIC model. *International Journal of Testing, 12*(1), 78-99. https://doi.org/10.1080/15305058.2011.620724

Muthén, L., & Muthén, B. (2012). *Mplus user's guide* (7th ed.). Muthén and Muthén.

Ojimba, D. P. (2012). Strategies for teaching and sustaining mathematics as an indispensable tool for technological development in Nigeria. *Journal of Mathematical Sciences, 3*, 23-35.

Ong, Y. M., Williams, J., & Lamprianou, I. (2015). Exploring crossing differential item functioning by gender in mathematics assessment. *International Journal of Testing, 15*(4), 337-355. https://doi.org/10.1080/15305058.2015.1057639

Oshima, T. C., Raju, N. S., Flowers, C. P., & Slinde, J. A. (1998). Differential bundle functioning using the DFIT framework: Procedures for identifying possible sources of differential functioning. *Applied Measurement in Education, 11*(4), 353-369. https://doi.org/10.1207/s15324818ame1104_4

Perry, J. L., Nicholls, A. R., Clough, P. J., & Crust, L. (2015). Assessing model fit: Caveats and recommendations for confirmatory factor analysis and exploratory structural equation modeling. *Measurement in Physical Education and Exercise Science, 19*(1), 12-21. https://doi.org/10.1080/1091367X.2014.952370

Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment, 29*(4), 347–363. http://doi.org/10.1177/0734282911406661

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment, 29*(4), 304–321. http://doi.org/10.1177/0734282911406653

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*(2), 159-194. https://doi.org/10.1007/BF02294572

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4), 589-617. https://doi.org/10.1007/BF02294821

Tata, U. S., Abba, A., & Abdullahi, M. S. (2014). The causes of poor performance in mathematics among public senior secondary school students in Azare Metropolis of Bauchi State, Nigeria. *IOSR Journal of Research & Method in Education, 4*, 32-40.

Tate, R. (2002). Test dimensionality. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 181-211). Lawrence Erlbaum.

Tsigilis, N., Gregoriadis, A., Grammatikopoulos, V., & Zachopoulou, E. (2018). Applying exploratory structural equation modeling to examine the student-teacher relationship scale in representative Greek sample. *Frontiers in Psychology, 9,* 733. https://www.frontiersin.org/article/10.3389/fpsyg.2018.00733

Umameh, M. A. (2011). *A survey of factors responsible for students' poor performance in mathematics in Senior Secondary School Certificate Examination (SSCE) in Idah Local Government Area of Kogi State, Nigeria* [Unpublished BSc(ED) thesis]. University of Benin.

Uwadie, I. (2012). *Federal government, teachers and parents battle students' under-performance*. Vanguard Newspaper. Retrieved September 23, 2022.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*(2), 103-118. https://doi.org/10.1207/s15324818ame0602_1

Wainer, H., Wang, X.-B., & Thissen, D. (1994). How well can we compare scores on test forms that are constructed by examinees choice? *Journal of Educational Measurement, 31*(3), 183-199. https://doi.org/10.1111/j.1745-3984.1994.tb00442.x

Walker, C. M., Zhang, B., Banks, K., & Cappaert, K. (2011). Establishing effect size guidelines for interpreting the results of differential bundling functioning analyses using SIBTEST. *Educational and Psychological Measurement, 72*(3), 415-434. https://doi.org/10.1177/0013164411422250

Wang, J., & Wang, X. (2012). *Structural equation modeling with Mplus methods and applications.* Wiley/Higher Education Press.