

# Test score equating of multiple-choice mathematics items: techniques from characteristic curve of modern psychometric theory

Musa Adekunle Ayanwale<sup>1</sup> 

Received: 24 April 2023 / Accepted: 22 August 2023

Published online: 31 August 2023

© The Author(s) 2023 [OPEN](#)

## Abstract

Examination scores obtained by students from the West African Examinations Council (WAEC), and National Business and Technical Examinations Board (NABTEB) may not be directly comparable due to differences in examination administration, item characteristics of the subject in question, and student abilities. For more accurate comparisons, scores from different assessments or examinations must be equated. In the literature, little is known about using Stocking-Lord and Haebara Item Response Theory (IRT) methods to equalize WAEC and NABTEB mathematics multiple-choice items. The study used a counterbalanced quantitative design for a single group. As a method for dealing with order effects, counterbalancing was used to assess 50 multiple-choice four-option items developed by WAEC and NABTEB for Grade 12 mathematics. The study selected 1210 Grade 12 students who were divided into groups A (647), and B (563). The IRT equating methods were employed to compare the two test forms based on the obtained data. The findings indicate that WAEC and NABTEB have similar construct domains, though WAEC is more difficult, and discriminates better. Further, when placed on the same scale, both tests are comparable. These findings have implications for curriculum redesign, academic performance assessment, and policymaking in the education sector, as well as refuting public perceptions of the credibility and value of these certificates awarded by the testing organizations in Nigeria.

**Keywords** Test equating · West African Examinations Council · National Business and Technical Examinations Board · Haebara · Stocking-Lord · Supplementary item response theory · EquateIRT

## 1 Introduction

Assessments are crucial in education as they provide valuable insights into a learner's knowledge and skills. In Nigeria, examination bodies like West African Examinations Council (WAEC) and the National Business and Technical Examinations Board (NABTEB) conduct high-stakes school leaving certificate examinations, and candidates use these certificates for higher education enrollment. In contrast to some testing programs that avoid repeating the same test to ensure security and cater to diverse candidates [1, 2], Nigerian examination bodies like WAEC and NABTEB administer tests multiple times throughout the year. However, a notable difference lies in Nigeria's context, where test questions can be leaked or taken out of test centers, potentially compromising the integrity of the original skill assessment. To address this issue, it is essential for the development of test items to prioritize the use of content and specifications/test blueprints whenever feasible. However, concerns have been raised about the perceived differences in the quality and difficulty levels of their test items. Some argue that test items from one body may be better or more favorable than the other, leading to worries

✉ Musa Adekunle Ayanwale, [ayanwalea@uj.ac.za](mailto:ayanwalea@uj.ac.za) | <sup>1</sup>Department of Science and Technology Education, Faculty of Education, University of Johannesburg, Johannesburg 2006, South Africa.



about discrimination in certificate awards. The need for equating WAEC and NABTEB stems from the potential inequity and inconsistency in the assessment process between the two bodies, which can impact students' opportunities and academic outcomes. Conducting an equating study to compare the mathematics test items and scores between WAEC and NABTEB is essential to determine the true comparability of their assessments and address concerns about potential bias and unfairness in the education system.

While test scores are often used for critical decisions by various entities, the nature of the exams taken by the examinees is rarely considered [3, 4]. The expectation is that parallel tests, such as NABTEB test A and WAEC test B, should generate equal scores across examinees. Equating test scores is vital for ensuring the validity, reliability, objectivity, and fairness of the results, allowing for interoperability between different types of tests and across different testing years. Previous research has established the comparability of scores from exams conducted by WAEC, NECO, or regional states in Nigeria [5–9] using traditional linear equating. However, this study seeks to fill both evidence and methodological gaps by standardizing WAEC and NABTEB mathematics items on a broad scale and analyzing them in-depth using IRT-based methods of Stocking–Lord and Haebara. By doing so, the ongoing debates and perceptions about the credibility of test items and certificates issued by these examination bodies can be resolved. Ultimately, this study holds great importance for an international audience as it addresses a crucial issue in educational assessment that extends beyond the borders of Nigeria. The study's focus on comparing mathematics test items and scores between two prominent examination bodies, WAEC and NABTEB, carries implications that resonate with educational systems worldwide. One of the key aspects that make this study relevant globally is its investigation of examination integrity. The study delves into the potential differences in test items and the comparability of scores between different examination bodies. Ensuring the integrity and fairness of examinations is a universal concern, as it directly impacts the credibility and reliability of educational qualifications in any country.

Furthermore, the findings of this study can significantly influence educational policies and standardization efforts at an international level. Policymakers and organizations involved in educational assessment can draw valuable insights from this study to enhance assessment practices and ensure equitable opportunities for students in diverse educational contexts. The study's implications for university admissions are also of broad interest. The findings can inform admissions processes in various countries, particularly those with multiple examination bodies. If test scores from different examination bodies are not comparable, it raises questions about the validity of using these scores as a basis for admission decisions in higher education institutions. The study's focus on fairness and equity in certificate awards based on the examination body is another aspect that appeals to an international audience. This issue is not unique to Nigeria, as other countries may also have concerns about potential discrimination in awarding certificates based on the examination body a student has taken. Additionally, the study's application of Item Response Theory (IRT)-based equating methods has broader implications beyond Nigeria's context. These methods have relevance for international assessments and standard-setting practices, making the study valuable for educational researchers and professionals around the world. Understanding the comparability of test scores is critical for ensuring the validity and accuracy of educational assessments, and this study contributes valuable insights that can be relevant to educational systems globally. Researchers worldwide can build upon this work to advance assessment methodologies and practices in their respective contexts.

In this paper, the following sections are presented. First, relevant and related literature including the underpinning theory, test score equating and its assumptions, score equating techniques equating designs, and IRT-equating methods are discussed in Sect. 2. The methodology adopted in the study is presented in Sect. 3, and Sect. 4 focuses on the analysis of data collected, and presentation of the results. In the final section, the author discusses the findings and highlights the implications for educational policy, and concludes by outlining limitations and future research prospects.

## 2 Literature review

### 2.1 The study underpinning theory

In test equating, Classical Test Theory (CTT), and Item Response Theory (IRT) are commonly used. CTT equating involves comparing observed or true scores at the examination level where the characteristics of score distributions are assumed to be equal for a specified group of examinees. CTT does not however provide assumptions of equality and invariance [10–13]. Various methods within CTT are employed for equating, such as linear equating, equipercentile equating, and several others. Contrarily, IRT methods, which form the basis of this study [14] model examinees' ability and item-level characteristics such as difficulty, discrimination, and guessing. IRT equating places estimates of item parameters on two

tests using the same scale, instead of relying on the distributions of total test scores. IRT likewise uses mathematical functions to describe the probability of correctly answering an item depends on the latent ability of the respondent [15–18]. The equating process in IRT involves estimating examinee and item parameters separately, and an individual's ability is not affected by item difficulty or ease [19]. Data collected by IRT can be correlated with any design, ranging from the statistically strongest single-group design to the weakest design that administers only a handful of items to the same group of participants [20, 21].

Additionally, different IRT models have distinct numbers of item parameters that define an item's statistical characteristics. The one-parameter logistic model (Rasch model or 1PL) only uses item difficulty to describe an item's characteristics and assumes all items are equally discriminating. The two-parameter logistic model (2PL) describes items by their difficulty and discrimination, thus reflecting their relationship with ability. The three-parameter logistic model (3PL) further includes a guessing parameter to account for examinees with low proficiency [22]. Although a four-parameter logistic model (4PL) has been developed, this study used the 3PL model for dichotomous responses in the data collection process (see Eq. 1). The 3PL model was chosen as a result of model-data fit assessment. When compared to 1PL and 2PL models, the 3PL model showed a better fit based on various information indices such as AIC (Akaike Information Criterion), SABIC (Sample-Size Adjusted BIC), HQ (Hannan-Quinn), BIC (Bayesian Information Criterion), logLik (log-likelihood), as well as likelihood ratio tests (logLik), chi-square statistic ( $X^2$ ), degrees of freedom (df), and p-value. Thus, the 3PL model's superior fit and greater parsimony led to its selection as the preferred model for fitting the datasets of WAEC and NABTEB.

$$P(Y_j = 1/\theta) = P_j = C_j + 1(1 - C_j) \frac{\exp\{a_j(\theta - b_j)\}}{1 + \exp\{a_j(\theta - b_j)\}} \quad (1)$$

where  $Y_j$  is the response to item  $j$ ;  $\theta$  is the ability;  $a_j$  is the discrimination parameter;  $b_j$  is the difficulty parameter, and  $c_j$  is the guessing parameter. If  $c_j = 0$ , the model reduces to the 2PL model. If  $c_j = 0$  and  $a_j = 1$ , the model reduces to the Rasch model.

## 2.2 Score equating and its assumptions

Test equating involves creating comparable scores for multiple versions of a test, thus enabling those versions to be interchanged. Researchers have devoted decades to the examination of test equating and found that scores are comparable when more than one form of the test is used [23–25]. Maintaining the technical quality of testing programmes is crucial for ensuring the validity of assessment results. The interpretation of scores is supported by evidence and theory, and equating is a statistical process that helps establish the same substantive meaning of scores on different test forms [26, 27]. Test equating as described by various researchers including [26, 28–32] involves adjusting test scores to make them comparable across different tests by accounting for differences in test difficulty and other statistical properties. Another perspective is that test equating establishes relationships between raw scores of different tests as explained by [33]. Consequently, for examining bodies, when administering test items in multiple instances, and to multiple examinee groups; equating method is encouraged as a statistical procedure to overcome the overexposure of items that can threaten test security [34]. Using equating methods with less error would ensure fairness among the examinees. Also, [35] explain equating test scores as an attempt to translate scores from one test to another defensibly. The question is: how can you determine if a score of 75 is the same if half of the students see one set of items and the other half see another set? Could one be a little easier? When one conducts assessments in linear forms, or pilots a bank of computerized adaptive tests/Linear-on-the-fly tests; one will likely use more than one test form, which requires test equating. As the result of equating test scores, the primary focus is to make them comparable, exchangeable, and interchangeable.

To achieve interchangeability and comparability between test scores, researchers suggest a five-step equating process [33, 36]. This involves ensuring that the tests measure the same construct, have equal reliability, use a symmetrical equating transformation, exhibit group invariance, and employ a common equating function [37, 38]. Equating instruments with significantly different reliabilities is discouraged [39]. Although achieving complete group invariance and population invariance is challenging [40], they can be evaluated through empirical assessments [41–44]. The equating requirements have been subject to various critiques, and a consensus on which ones are crucial remains elusive [45, 46]. Nonetheless, the primary objective of equating is to make scores on different test forms comparable, with an emphasis on aligning the tests to the same construct. In this study, the focus was on equating the tests using the same construct, as both tests were designed to measure the same content in the senior secondary school mathematics curriculum.

### 2.3 Score equating techniques

Equating scores from different test forms/testing programs can be accomplished using several techniques, processes, or methodologies. As noted by [47, 48], equating can be classified into three types: vertical, score linking, and horizontal. As a method of comparing student scores across tests of different or multiple levels, vertical equating is also known as across-grade equating or scaling. It compares the content and difficulty of the tests across grade levels in the same construct, thereby summarizing student progress over time ([49], p. 50). Often, equating is used as a developmental scale. For instance, mathematics tests for grade 11 and grade 9 are compared. In these tests, mathematical skill is the focus, even though their content differs. The ability to do mathematics should steadily increase every year as students improve. In vertical equating, two different groups of examinees are compared, so it is complex conceptually. Moreover, two tests given to the same group of students such as those in grade 8 and grade 9 are likely to be easier in grade 9, thus resulting in a ceiling effect that will not provide accurate performance information at the second administration. The horizontal equating technique, also known as within-grade equating, compares the scores of similar tasks among students at the same level, topic, and population [49–52]. It is used when students re-take exams on different forms, all of which are equated to achieve comparable scores. This method is straightforward, comparing groups of examinees with the same ability level using different tests based on the same content and difficulty range, such as those used in WAEC and NABTEB Mathematics test items. There are two types of equating: pre-equating and post-equating. Pre-equating converts raw scores into scaled scores before the operational test is administered, using data from field tests analyzed with statistical procedures [53]. Post-equating, on the other hand, adjusts operational test data for difficulty after administration and is considered more accurate. Pre-equating helps with score reporting, quality control, and assessment flexibility, but it may lead to motivation issues if field test items are presented separately. Post-equating is more accurate but requires sufficient time. When time is limited, the equating process may be compromised, affecting the reporting process and quality control.

### 2.4 Score equating designs

Various testing equating designs include single-group designs, counterbalance single-group designs, equivalent-group designs, and non-equivalent anchor test designs (NEAT). In a single-group design, two testing instruments (X and Y) are administered to one group of participants to estimate item parameters. The drawback of this single-group design is possible fatigue or familiarity effects. Counterbalanced designs involve two sample groups taking X and Y in different orders to control for order effects. The advantage is accurate results with a small sample size [43]. Equivalent group designs administer separate test forms to the same group, but equivalence assumptions can be problematic [54]. In non-equivalent group designs, different populations take the test forms [1]. Assert that NEAT is widely used for administrative flexibility but can be challenging due to non-equivalence of test forms [55], equating based on different sets of common items can yield different results when equating two test forms. Upon discussing the results later [56], reported that when the groups were similar in ability, the anchor tests yielded similar equating results. The author further reported that when the groups differed in their ability level, the anchor tests yielded very disparate equating results. Consequently, the anchor test must be carefully chosen [56]. Short tests are less reliable than long tests. To reflect group differences accurately and reliably, anchor tests need to be long enough. There is a less random equating error when there are more common items [57–59]. More importantly, common person and item equating are also distinctions in the methods of equating. Common person equating involves administering two tests to a group of people. By using a linear transformation, we calculate the mean and standard deviation of the scale locations for the two groups. Due to the same person taking both tests, the study proposes a common person equating. Notably, an anchor test embeds a set of common items into two different tests, and thereby resulting in a common item equating. Common items are equated by their mean location. Four steps are involved in equating, and they are: data collection, defining an operational equating transformation, selecting an estimation method, and evaluating the results.

### 2.5 IRT equating methods

There are four steps involved in the equating process through IRT. First, the ability and item parameters are derived from the equating design or data collection design. Second, ability and item parameters are estimated or calibrated either simultaneously or individually for each form of the test. Ability parameters can therefore be calibrated simultaneously

or separately [60]. During concurrent calibration, item parameters of two forms of a test are calibrated simultaneously, whereas separately calibrated item parameters are calibrated separately through the calibration software for each form. Lastly, item parameters across different test administrations must be equated on a common scale since equating cannot take place without common item parameters. In IRT, scaling coefficients can be obtained from different calibrations of ability and item parameters on a common metric using various methods. Among them are Haebara and Stocking-Lord (characteristics curves method), mean/mean, and mean/sigma (moment methods). Haebara and Stocking-Lord take into account both difficulty and discrimination parameters of the test [61, 62]. A transformation constant is identified first in this model. A comparison of different tests or item characteristic curves is thereafter performed, and then investigate why the differences decrease. Consequently, a comparison of Haebara and Stocking-Lord (characteristics curves) was conducted in this study to determine the comparability of mathematics items administered by WAEC and NABTEB in Nigeria. By summing the squares of the differences in each item's characteristic curve, the Haebara method estimates the differences between item characteristic curves [62, 63]. Here is the mathematical expression for this method:

A: Curve of the equating equation

B: Constant of the equating equation

$P_{ij}(\theta_{ji}, a_{ji}, b_{ji}, c_{ji})$ : Item characteristic function.

$P_{ij}\left(\theta_{ji}, \frac{a_{ji}}{A}, Ab_{ji} + B, c_{ij}\right)$ : Equated item characteristic function

$$Haediff(\theta_i) = \sum_{j:v} \left[ P_{ij}(\theta_{ji}, a_{ji}, b_{ji}, c_{ji}) - P_{ij}\left(\theta_{ji}, \frac{a_{ji}}{A}, Ab_{ji} + B, c_{ij}\right) \right]^2 \quad (2)$$

$$Hae_{crit} = \sum_i Haediff(\theta_i) \quad (3)$$

Moreover, by taking the square of the sums of differences between each item's characteristic curve, Stocking Lord (1983) calculates the difference between them. Below is a mathematical expression of this method:

A: Curve of the equating equation.

B: Constant of the equating equation.

$P_{ij}(\theta_{ji}, a_{ji}, b_{ji}, c_{ji})$ : Item characteristic function.

$P_{ij}(\theta_{ji}, Ab_{ji} + B, c_{ij})$ : Equated item characteristic function

$$Stldiff = \left[ \sum_{j:v} P_{ij}(\theta_{ji}, a_{ji}, b_{ji}, c_{ji}) - P_{ij}\left(\theta_{ji}, \frac{a_{ji}}{A}, Ab_{ji} + B, c_{ij}\right) \right]^2 \quad (4)$$

$$Stl_{crit} = \sum_i Stldiff(\theta_i) \quad (5)$$

## 2.6 Previous studies

Numerous studies (e.g., [2, 46, 53, 57, 62, 64–70]) have investigated comparability of test-takers scores from different examination bodies using Item Response Theory (IRT) equating methods, including Haebara, Stocking-Lord, mean-mean, mean-sigma, and concurrent calibration. These studies have shown that certain equating methods perform better than others in various contexts of examination test items. For example, [46] compared the Haebara and Stocking-Lord methods using data from the 2018 National Examination Administration, and the Haebara method was found to have higher mean-sigma values, suggesting potential improvements in discrimination and difficulty level. Another study [65] employed the Rasch Model to equate mathematics test scores, demonstrating the effectiveness of IRT-based methods in this context. Similarly [66], compared various equating methods and found that both Haebara and Stocking-Lord methods yielded comparable results in terms of test score comparability for mathematics assessments. In the context of WAEC mathematics test items, [70] compared CTT and IRT equating methods and found that IRT's mean-sigma method outperformed with smaller errors. Additionally [2], explored linking and concurrent calibration methods using mixed-format tests under IRT for non-equivalent groups with common-item designs, and the concurrent calibration method

generally performed better, recovering item parameters accurately and generating more precise estimated scores. Furthermore [57], identified that IRT methods produced smaller equating errors compared to CTT methods in the context of SAT-Verbal tests. Studies like [71] and [72] investigated the comparability of mathematics examination items between different examining bodies and found varying degrees of equivalence. Despite these valuable insights, there has been no specific study in Nigeria focusing on the comparison of mathematics test items between WAEC and NABTEB using the Stocking-Lord and Haebara methods of IRT. This present study aims to fill this research gap and contribute to the existing body of knowledge.

### 3 Materials and method

#### 3.1 Philosophical lens

The paradigm tells the reader how to interpret the research results based on the collected data [73]. It is useful for the researcher to select a paradigm, because without it, he/she cannot focus on a particular philosophical knowledge and evaluate other possibilities. This study is based on a post-positivist research philosophy. In postpositivism, factual knowledge, which includes measurement is considered trustworthy. Ultimately, this study is shaped and driven by the postpositivist paradigm's beliefs, convictions, expectations, and values [73]. Additionally, the study employed a cross-sectional quantitative method of a single group with a counterbalance design that requires each examinee to take both tests in a counterbalanced fashion. Since the same examinees write both tests and have the capacity to boost the validity of the dataset used in the study, the equating design is regarded as the strongest statistical design since it uses examinees with presumed equal ability, which also meets Lord's equity requirement.

#### 3.2 Participants

Detailed demographic information about Grade 12 students in government-owned schools who voluntarily participated in this study is shown in Table 1. In addition, consent was sought from students and administrators before participating in the study in order to maintain ethical standards. The two test forms attracted 1,300 responses in total, and 1,210 responses were useful in the analysis. A greater percentage of male students responded to the test forms, thus representing 51.7% of the population. When data was being collected, the age range of the students was within 17–19 years, suggesting the expected age group for the student to sit for terminal examinations conducted by public examining bodies. Even though students across different locations in educational district II of Lagos State, Nigeria were sampled, more of the students in this study (i.e., 45.1%) were from schools within Ikorodu area. Although 44.5% of the students were of science extraction, the sample included other students with, commercial, and humanities specializations. As part of the study design, samples were also classified into group A (with 647 students), and group B (with 563 students) respectively.

**Table 1** Demographic profile of the participants

Variable	Categories	Frequency	Percentages
Gender	Male	625	51.7
	Female	585	48.3
Age group	14–16	368	30.4
	17–19	695	57.5
	20–22	147	12.1
School location	Kosofe	439	36.3
	Somolu	225	18.6
	Ikorodu	546	45.1
Specialization	Science	539	44.5
	Commercial	410	33.9
	Humanities	261	21.6

### 3.3 Measures

WAEC is considered a large-scale exam as it involves a substantial number of candidates taking the test. Furthermore, it is administered to a wide-ranging population across various West African countries, with Nigeria being one of them. All Grade 12 students are required to sit for the mathematics test, one of the cross-cutting subjects. Students use the results of this test to apply to higher education institutions (HEIs). It is therefore considered high-stakes, and the items on the test must sufficiently measure traits in a valid and reliable manner. The test comprises 50 items from a wide range of content domains, each with one correct answer, and three dummy options. Several stages of development and validation have been undertaken to standardize the test. By shading an optical mark reader (OMR), the participants scored 1 for each correct answer and 0 for each incorrect answer. Also, there is often a correction for guessing, whereby test-takers are typically awarded points for correct answers but receive a penalty or no points for incorrect answers. The aim is to discourage random guessing and encourage test-takers to answer only when they are reasonably confident in their responses. The specific scoring scheme may vary, but the idea is to prevent several wrong answers from outweighing correct ones and affecting the overall score negatively.

Furthermore, NABTEB is another Nigerian public examination body entrusted with the responsibility of conducting technical and business innovation certificate examinations, which WAEC had previously conducted [74]. Similarly, NABTEB examinations are held twice a year in May/June and November/December. In this study, their May/June mathematics test items, which comprised 50 items with four options, letters A through D were utilized. All test items were dichotomously scored (0 for each incorrect answer, and 1 for each correct answer). A raw score of 50 was the maximum possible, and a score of 0 was the minimum. The items were taken from their high school mathematics curriculum, which had almost the same content as WAEC's. Using the two test forms, this study examined whether the scores are comparable when put on the same continuum or scale. Subject experts have validated the instruments used regarding their content. Based on the loading factor and construct validity analysis, the loading factor and CVI were greater than 0.50. At the same time, the IRT empirical reliability coefficients returned 0.89 and 0.86 respectively. The instruments can therefore be used for equating.

### 3.4 Data analysis

Through different R packages, Haebara and stocking-lord equating methods were implemented to analyze the obtained data (see Appendix-session 1 for codes on the importation of data into the Rstudio environment for data analysis). First, the assumption of similar construct for the two forms was tested using *sirt* package in R software version 4.0.1 [75, 76]. In this package, the confirmatory detect (conf detect) function is used to establish Stout's test of essential unidimensionality assessed by the DETECT index [77–79]. Under a confirmatory specification of item clusters, this function computes the dimensionality evaluation to enumerate contributing traits (DETECT) statistic for dichotomous item responses, and the polyDETECT statistic for polytomous item responses [77, 80–83]. Noticeably, DETECT produces indices, including dimensionality evaluation to enumerate contributing traits, approximate simple structure index (ASSI), and approximate simple structure index ratio index (RATIO) [82]. The option unweighted means that all conditional covariances of item pairs are equally weighted, and weighted means that the sample size of item pairs weights these covariances. The following classification scheme is used to determine the dimensionality of the test [83, 84].

Strong multidimensionality     $\text{DETECT} > 1.00$

Moderate multidimensionality     $0.40 < \text{DETECT} < 1.00$

Weak multidimensionality     $0.20 < \text{DETECT} < 0.40$

Essential unidimensionality     $\text{DETECT} < 0.20$

Maximum value under simple structure     $\text{ASSI} = 1 \text{ RATIO} = 1$

Essential deviation from unidimensionality     $\text{ASSI} > 0.25 \text{ RATIO} > 0.36$

## Essential unidimensionality ASSI &lt; 0.25 RATIO &lt; 0.36

Also, scores from the two test forms were transformed through R packages such as multidimensional item response theory (*mirt*) [14], which gave the user the option of modeling response data using several IRT models. Based on the *mirt* package, IRT objects could be read into the *equateIRT* package [84]. The use of IRT models [85–88] is widespread in the present day for the analysis and scoring of tests. It is natural to use IRT equating because many testing programmes assemble their tests using IRT [19, 85, 89]. When fitting the dichotomous data matrix  $k_x \times J_x$  with an IRT model, the parameter estimates,  $\hat{\theta}$  and  $\hat{\omega}$  for both persons and items will be obtained. A test score based on an IRT model is typically a prediction of the ability parameter ( $\hat{\theta}$ ) known as an IRT score, instead of the sum score  $X_i \in x$ . A transformation function is needed for a test form X and a test form Y to have equivalence IRT scores under the IRT setting. Depending on the equating design used, IRT scales will be transformed differently. When estimating single-group counterbalance designs, the abilities are not a concern; no additional transformations of scales are needed if the mean and variance of the ability distribution are assumed [32]. Kolen et al., [26] Show that IRT parameter estimates from two calibrations of different test forms should be on the same scale, which is done by using IRT parameter linking [90].

The functions *modIRT()*, and *direct()* are therefore also included in *equateIRT* to implement IRT parameter linking/equating. When *modIRT()* is invoked, the *coef* argument accepts a matrix of item parameter estimates. Aside from estimating the standard errors for estimated equating coefficients, *equateIRT* can also estimate the covariance matrix of item parameter estimates. In summary, equating scores on a new test form 1 to the scale of test form 2 can be achieved in two steps. A linear transformation is used to rescale the item parameters of form A onto the scale of form B using IRT methods such as the item characteristic curve [61, 91, 92]. Because the IRT item parameters of the two test forms are the same, the scores on these two tests are also the same since they are direct functions of the IRT item parameters. When multiple new test forms (e.g. X1, X2,..., Xn) need to be equated to a base form, score equating is an efficient approach. Succinctly, test scores are equated by using the 3-PL IRT model for multiple-choice mathematics questions. To ensure that test scores are comparable between different forms or administrations of the test, item difficulty, and person ability parameters are estimated using item response data. In addition, the maximum likelihood estimation procedure was used for estimating the coefficients and calibrating the items.

## 4 Results

To check the major requirement of the similar construct as recommended by [38, 72, 93], the responses of the examinees to the two mathematics tests (prepared, administered, and assessed by WAEC and NABTEB) were subjected to Stout's test of essential unidimensionality test implemented in the *sirt* package of R Language and environment for statistical computing (see Appendix-session 2 which describes R codes for establishing a Stout test of essential unidimensionality of the two test forms prior to the equating process). Table 2 present the result.

Table 2 shows that the two test forms of mathematics multiple-choice items were essentially unidimensional; the form1 had maximum DETECT value = -0.304 (< 0.20), ASSI = -0.476 (< 0.25), and RATIO = -0.678 (< 0.36), and form2 had maximum DETECT value = -0.303 (< 0.20), ASSI = -0.469 (< 0.25) and RATIO = -0.671 (< 0.36), respectively. The assumption of unidimensionality was therefore not rejected for the two tests. This result shows that one dominant dimension accounted for the variation observed in students' responses to the two tests of mathematics multiple-choice items tests. The mathematics multiple-choice items tests thereby fulfilled the similar construct requirement of conducting tests equating and unidimensionality assumption of item response theory. Table 3 presents 3PLM item parameters for the two test forms.

**Table 2** Dimensionality assessment of the two tests

	Form 1*		Form 2*	
	Unweighted	Weighted	Unweighted	Weighted
DETECT	-0.304	-0.304	-0.303	-0.303
ASSI	-0.476	-0.476	-0.469	-0.469
RATIO	-0.678	-0.476	-0.671	-0.671

\*Form1-NABTEB; Form 2-WAEC



**Table 3** Item parameters for the two test forms

Items	Form 1*			Form 2*		
	a	b	g	a	b	g
I01	1.17	0.08	0.39	0.89	-1.17	0.01
I02	0.91	0.46	0.29	1.21	-1.14	0.12
I03	1.03	2.05	0.01	0.92	-1.55	0.00
I04	1.19	0.83	0.08	1.26	-1.07	0.03
I05	1.61	-2.09	0.29	1.05	-0.34	0.01
I06	1.20	1.38	0.13	1.09	0.23	0.00
I07	1.02	1.52	0.00	1.30	0.89	0.01
I08	1.11	-0.64	0.06	0.99	-0.72	0.00
I09	1.05	0.59	0.00	0.93	1.52	0.01
I10	0.99	1.55	0.00	1.27	0.10	0.00
I11	0.93	0.84	0.01	1.40	-0.49	0.07
I12	1.05	0.58	0.00	1.25	-0.83	0.06
I13	0.92	0.42	0.04	1.11	-0.44	0.01
I14	1.15	-1.71	0.13	1.01	0.45	0.01
I15	1.27	-0.59	0.07	1.03	0.59	0.07
I16	1.05	1.02	0.04	1.21	-0.22	0.05
I17	1.16	0.18	0.09	1.15	0.55	0.13
I18	0.87	0.00	0.02	1.02	-1.86	0.00
I19	1.08	-0.85	0.01	0.93	-1.72	0.06
I20	0.92	-0.68	0.00	1.28	-0.72	0.26
I21	1.15	-0.50	0.08	1.25	0.07	0.16
I22	1.07	-0.03	0.01	1.17	-0.30	0.18
I23	1.06	0.44	0.04	0.88	0.59	0.00
I24	1.08	-0.08	0.00	0.98	-0.73	0.00
I25	1.24	1.85	0.11	1.01	0.90	0.00
I26	1.11	-1.05	0.00	1.06	-0.64	0.31
I27	1.02	1.25	0.00	1.10	1.78	0.00
I28	1.02	-0.74	0.00	1.67	0.80	0.27
I29	0.91	-0.32	0.00	1.17	-0.34	0.07
I30	1.52	-1.01	0.03	1.10	1.50	0.00
I31	1.00	1.07	0.01	1.21	0.85	0.00
I32	0.97	-0.64	0.00	1.26	1.74	0.27
I33	1.02	-1.90	0.03	0.93	1.18	0.04
I34	1.07	-0.15	0.03	1.03	-0.05	0.07
I35	1.11	1.66	0.05	1.10	0.84	0.02
I36	1.12	-0.68	0.02	1.14	-1.40	0.05
I37	1.00	0.65	0.00	1.28	1.18	0.05
I38	1.05	1.62	0.13	0.95	-1.28	0.00
I39	1.13	0.96	0.00	0.98	1.19	0.02
I40	1.19	-1.03	0.00	0.97	1.20	0.01
I41	1.35	1.14	0.01	0.96	1.62	0.00
I42	1.15	-1.30	0.04	0.95	0.60	0.01
I43	0.97	-0.98	0.00	1.16	1.41	0.25
I44	1.09	1.56	0.00	0.97	1.51	0.01
I45	1.04	-0.35	0.00	1.14	2.00	0.00
I46	1.02	-1.17	0.01	1.08	0.25	0.03
I47	0.96	0.03	0.00	1.29	-1.30	0.07
I48	1.04	0.64	0.00	0.92	-1.10	0.00
I49	1.41	1.06	0.32	1.37	-0.49	0.07
I50	1.16	-1.48	0.00	1.20	1.67	0.14
<b>Mean</b>	<b>1.09</b>	<b>0.11</b>	<b>0.05</b>	<b>1.11</b>	<b>0.15</b>	<b>0.06</b>
<b>SD</b>	<b>0.15</b>	<b>1.07</b>	<b>0.09</b>	<b>0.16</b>	<b>1.08</b>	<b>0.08</b>

**Table 3** (continued)

The bold indicates the means and standard deviations for the item parameters for form 1- NABTEB and form 2- WAEC, respectively.  
\*Form1-NABTEB; Form 2-WAEC

**Table 4** Characteristics curve equating coefficient of the two test forms

	Stocking-Lord estimate	Haebara estimate	StdErr
A	1.028	0.965	NA
B	0.085	0.227	NA

Table 3 displays parameters generated using 3PL implemented in the mirt package of R language (see detailed description of the R codes for establishing item calibrations (3PL) of the two test forms before equating process in [Appendix-session 3](#)). In the following estimates,  $a$  represents discrimination/slope,  $b$  represents difficulty/threshold, and  $g$  represents the guessing parameter. Form 2 discriminated between examinees who knew the subject material, and those who did not ( $M = 1.11$ ,  $SD = 0.16$ ) more strongly than Form 1 ( $M = 1.09$ ,  $SD = 0.15$ ). As a result, test items from form 2 distinguish better between examinees with low and high abilities. The difficulty indices indicate how easy or difficult the two test forms were for the examinees. In [\[92, 94\]](#), the authors recommend that easier items have lower difficulty indices (negative values) and that very easy items have values less than  $-2$ , while harder items have higher indices (positive values) and very hard items have values greater than  $+2$ . Accordingly, form 2 test items were moderately difficult ( $M = 0.15$ ,  $SD = 1.08$ ) as compared to form 1 test items ( $M = 0.11$ ,  $SD = 1.07$ ) in terms of difficulty. The guessing parameter is also considered unacceptable if the test item falls outside this range of  $c > 0.35$  [\[95\]](#). On average, both test forms ( $M = 0.05$ ,  $SD = 0.09$ ,  $M = 0.06$ ,  $SD = 0.08$ ) were not vulnerable to guessing. Using the equateIRT package implemented in R language, the equating coefficients were obtained for both test forms using stocking-lord and Haebara equating methods as presented in [Tables 4 and 5](#). Meanwhile, the R codes for establishing the equivalence and comparability of the two test forms using the Stocking-lord and Haebara IRT methods can be found in [Appendix-session 4](#).

The Stocking-Lord procedure and Haebara method were compared for computing equating coefficients, and it was discovered that they required different levels of arithmetic complexity. The Stocking-Lord procedure utilized a quadratic loss function to minimize the equating coefficients based on the test characteristic curve, while Haebara's method involved calculating the probability of correct response for each item in the test using item parameter estimates from test calibration, and then computing a similar vector for the same theta value. The slope and intercept coefficients were calculated using a weighted least squares procedure to minimize the average squared difference between the ability distribution and the test item parameter estimates. [Table 4](#) displays how the equating coefficients produced by the two methods yield similar estimates of item and person parameters when used for common equating operations. The focus of the comparison is on the agreement between the two methods, rather than the goodness of fit of the transformed values to the underlying parameters. The StdErr column in the output however shows NA, as the researcher did not provide the covariance matrix for the item parameter estimates.

[Table 5](#) in the provided data displays equated scores for two forms of a test using two different methods, namely Stocking-Lord and Haebara to allow for comparison. The table has four columns with the second and fourth columns representing the possible observed scores ranging from 0 to 50 on the two test forms. The column labeled 'form1.as.form2' displays the equated scores for form 1 using both methods. For instance, when using Stocking-Lord, a score of 0 on form 2 is equivalent to  $-0.11$  on form 1, a score of 1 is equivalent to  $0.87$ , and so on. Similarly, using Haebara, a score of 0 on form 2 is equivalent to  $-0.49$  on form 1, a score of 1 is equivalent to  $0.38$ , and so on. The equated scores for both methods were relatively similar, with an average (mean) of  $24.98$  and a standard deviation (SD) of  $14.91$  for Stocking-Lord; and an average of  $25.07$  and an SD of  $17.87$  for Haebara. Relatedly, if a student scored 50 on form 2, they would be expected to earn a score of  $50.06$  on form 1 using Stocking-Lord and  $50.49$  using Haebara. To visually compare the two equating methods, a plot ([Fig. 1](#)) displays the equated scores from the test forms against the scale scores.

[Figure 1](#) consistently shows that equating form 1 and form 2 using the IRT methods results in a much stronger claim that the scores can be correctly interpreted. Due to these equated scores, the two public examining bodies are measuring very similar knowledge and skills on the test forms.

**Table 5** Equating methods on the two test forms

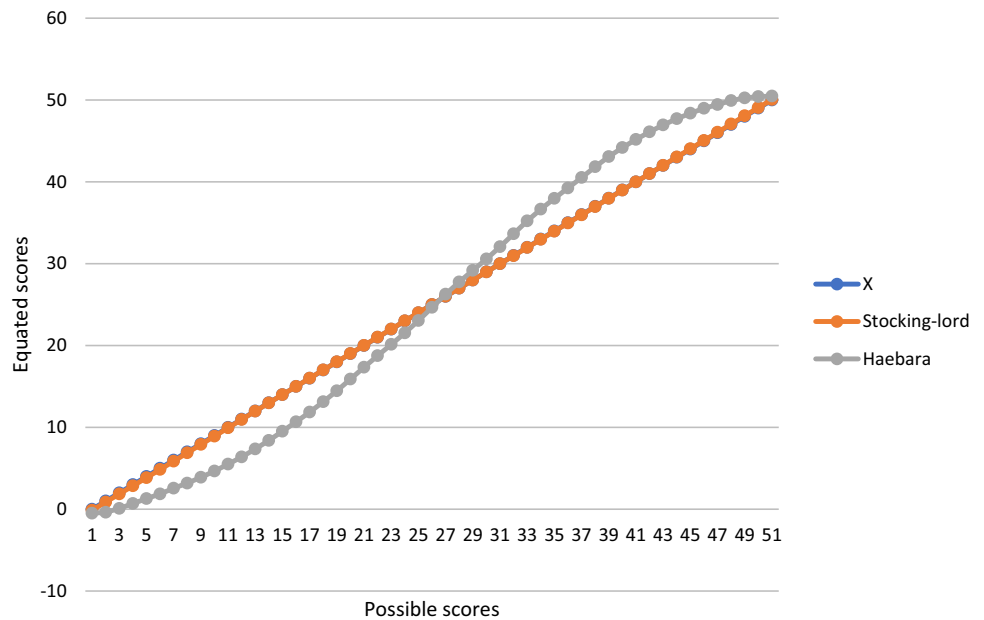
Theta	Stocking-Lord		Haebara	
	Form2	Form1.as.Form2	Form2	Form1.as.Form2
-4.0	0	-0.11	0	-0.49
-3.3	1	0.87	1	-0.38
-2.9	2	1.86	2	0.09
-2.5	3	2.85	3	0.69
-2.3	4	3.85	4	1.30
-2.0	5	4.86	5	1.87
-1.8	6	5.87	6	2.56
-1.7	7	6.89	7	3.19
-1.5	8	7.91	8	3.90
-1.4	9	8.92	9	4.67
-1.2	10	9.94	10	5.51
-1.1	11	10.96	11	6.39
-1.0	12	11.98	12	7.36
-0.8	13	12.99	13	8.41
-0.7	14	14.00	14	9.53
-0.6	15	15.01	15	10.68
-0.5	16	16.01	16	11.87
-0.4	17	17.02	17	13.13
-0.3	18	18.02	18	14.47
-0.2	19	19.02	19	15.90
-0.1	20	20.02	20	17.35
0.0	21	21.02	21	18.76
0.1	22	22.02	22	20.14
0.2	23	23.02	23	21.54
0.3	24	24.02	24	23.05
0.4	25	25.01	25	24.67
0.5	26	26.01	26	26.28
0.6	27	27.00	27	27.77
0.7	28	28.00	28	29.19
0.8	29	28.99	29	30.59
0.9	30	29.99	30	32.08
1.0	31	30.98	31	33.66
1.2	32	31.98	32	35.23
1.3	33	32.98	33	36.67
1.4	34	33.97	34	37.98
1.5	35	34.97	35	39.25
1.7	36	35.97	36	40.53
1.8	37	36.98	37	41.85
2.0	38	37.98	38	43.09
2.1	39	38.99	39	44.19
2.3	40	39.99	40	45.18
2.5	41	41.01	41	46.11
2.8	42	42.02	42	46.96
3.1	43	43.03	43	47.72
3.3	44	44.04	44	48.38
3.5	45	45.06	45	48.99
3.6	46	46.07	46	49.45
3.7	47	47.07	47	49.94
3.8	48	48.08	48	50.25
3.9	49	49.07	49	50.42
4.0	50	50.06	50	50.49

**Table 5** (continued)

Theta	Stocking-Lord		Haebara	
	Form2	Form1.as.Form2	Form2	Form1.as.Form2
	<b>Mean</b>	<b>24.98</b>	<b>Mean</b>	<b>25.07</b>
	<b>SD</b>	<b>14.91</b>	<b>SD</b>	<b>17.87</b>

The bold indicates the means and standard deviations for equating form 1- NABTEB as form 2- WAEC using Stocking-Lord and Haebara procedure

**Fig. 1** Equated scores of the test forms



## 5 Discussion and conclusion

As a result of continuous diverse public perceptions over the superiority of two public examining bodies in Nigeria pertaining to placement, making decisions, awarding scholarships, and accepting admissions to universities and colleges; the present study attempted to compare mathematics test items of both examining bodies to determine their score comparability. By using the DETECT confirmatory of the Stout test of essential unidimensionality implemented in R, it was possible to establish the most important test equating assumption of the same construct. This study has demonstrated that both test forms comply with the single construct that explains the performance of the two tests examined. The study finding is in agreement with [5, 70] that WAEC diets of June and November had the same construct specification. As the two examination bodies are using the same mathematics curriculum/syllabus to develop their test items, the two tests measure the same content and cognitive processes. The two examination bodies support the same inferences about what students know and can do, thus, a strong claim has been established in this study to correct varied public perceptions/opinions on the disparity and quality of certificates awarded by these two bodies. Also, based on the calibration of the two forms using 3PL, form 2 was found to be moderately difficult compared to form 1, and therefore useful to differentiate between examinees with low abilities and those with high abilities more effectively. No strong evidence was established that low-ability students can guess the correct answer to the items in the two forms randomly, as indicated by the guessing parameters of the two forms.

Further, the study utilized two different equating procedures, namely Stocking-Lord method and Haebara method to ensure comparability and equivalence of the two test forms despite having different underlying arithmetic characteristics. The results showed a strong correlation between the equating scores obtained from both methods, and thereby indicating a reasonable level of agreement between them. This finding is consistent with previous studies [26, 46, 65, 66, 71] that have demonstrated the effectiveness of using IRT-based equating methods to compare test scores across different forms or administrations of a test. In line with [46], this study confirms that the Haebara

method exhibits higher mean-sigma values, suggesting potential improvements in discrimination and difficulty level. Similarly [65], demonstrated the effectiveness of IRT-based methods, which aligns with the use of the 3PL model in this research to equate mathematics test scores. The study concurs with the findings of [66], where both Haebara and Stocking-Lord methods yielded comparable results in terms of test score comparability. Studies like [26] and [71] that have investigated the comparability of examination items between different examining bodies resonate with the findings of this study. Furthermore, the study findings indicated that the mathematical test form from WAEC exhibited slightly higher difficulty and better discriminatory power among examinees compared to the mathematical test form from NABTEB, as evident from the estimated item parameters. These differences in item parameters could be attributed to specific characteristics of the test items or variations in the populations of examinees from the two examining bodies. For instance, the WAEC test may have been designed to assess a higher level of proficiency or administered to a more advanced group of examinees compared to the NABTEB test. This discovery highlights the significance of equating test scores to account for such differences in item parameters and ensure fair score interpretation across different test forms or examinee populations. Equating test scores becomes crucial to achieving score comparability and making valid comparisons across different test forms or examination bodies. Interestingly, it was somewhat surprising to note that despite the significant differences in the mathematical stringent nature of the two procedures, they produced very similar estimates of item and ability based on a given set of items. This observation underscores the robustness of equated metrics in producing comparable estimates, even when test forms differ significantly. In conclusion, the study demonstrates that the two test forms possess similar characteristics, if not the same. Therefore, there should be no doubt or discussion about the comparability of their results or the usability of their certificates without any reservations. Equating the scores between WAEC and NABTEB is crucial for ensuring fairness and equity in score interpretation, making the certificates from both examining bodies equally valid and reliable for various purposes.

## 6 Implications for educational assessment and policy

There are profound implications for educational policy and practice in Nigeria based on the study findings. Based on the EquateIRT package's equating results, test scores are comparable across different examination bodies, thus demonstrating that test scores can be used fairly and equally when making important decisions, including curriculum redesign, student placement, admissions, and promotions. By combining test scores, we can interpret and compare scores meaningfully across different test forms and examination bodies, and making it easier to make accurate and reliable decisions based on test scores. In the education system, this can have a significant impact on ensuring fairness, transparency, and accountability. Furthermore, this study contributes to the growing body of research on equating methods and their applicability in diverse educational contexts. Based on this study, the EquateIRT package has been found to be useful and effective in equating test scores among Nigerian educational assessment organizations. Research and practitioners can therefore use EquateIRT to equalize test scores in a variety of settings with its user-friendly and flexible design. More so, this study's implications extend beyond the Nigerian context, offering valuable lessons and insights to an international audience. It explores methods of equating, like Haebara and Stocking-Lord, which can guide countries facing challenges in comparing test scores. Ensuring fairness and comparability in test scores is a concern shared worldwide and educational systems can learn from this study's approach to accurately interpret scores. The study's use of IRT to assess test score comparability emphasizes the importance of employing methodologies to validate assessments. Policymakers and educational institutions around the world should consider the significance of equating methods in maintaining assessment practices and providing opportunities for students. Moreover, the findings from this study contribute to discussions on fairness and equity in education on a level. It is crucial to have fair test scores when making decisions about student performance, admissions and educational opportunities. By showcasing practices in equating this study provides guidance for countries to enhance their own assessment procedures and ensure precise interpretations of test scores. International collaboration in research is vital, as sharing knowledge and experiences regarding equating assessments can foster a comprehensive understanding of global test score comparability. In conclusion, while the study's primary focus is on Nigeria, its implications and messages are relevant to educational systems worldwide. By sharing these insights, this research has the potential to enhance assessment methods and foster equal educational opportunities beyond Nigeria's boundaries.

## 7 Limitations and future research

The sample size used in this study provides valuable insights into the research question, even though the scores obtained from the sampled Nigerian students are not representative of all Nigerian students. To further validate the findings, future research on this topic should consider a larger sample size that includes students from other geopolitical zones in Nigeria. Furthermore, future studies should examine other methods of data collection design and equating methods using Item Response Theory (IRT) to compare scores from other subjects administered by public testing agencies. In future studies, it will be necessary to establish the standard error of the equating coefficient, and take into account the differences between observed and true scores. Future findings would be more robust if these considerations are taken into account.

**Acknowledgements** The author acknowledges the support received from the principals and teachers of all the participating schools, and their students for seamless cooperation throughout the data collection phase of the study.

**Author contributions** MAA: conceptualization, writing—original draft, methodology, data curation, validation, formal analysis, visualization, formatting, references alignment, writing review & editing.

**Funding** This research received no external funding.

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

## Declarations

**Ethics approval and consent to participate** In accordance with the ethical guidelines of the Declaration of Helsinki, this study was approved by the institutional review board of the University of Johannesburg (Ethical approval number: Sem 2-2022-162). This study involved Grade 12 students from Lagos State's Educational District II. Participant's parents or guardians gave informed consent before participating in the study. The study participants were informed they could withdraw at any time without adverse consequences. Participants' data were stored securely and anonymously with unique identification codes. The study minimized risks to participants and ensured their safety. There was no harm or discomfort experienced by the participants, and their rights, welfare, and dignity were fully protected.

**Consent for publication** This study used human and informed consent was obtained from all subjects involved.

**Competing interests** The author declares no competing interest.

## Appendix

### Code lines used in Rstudio software

#### Session 1 – Data importation into R.

```
# get the working directory.  
getwd().  
# set the working directory.  
setwd("C:/Users/DELL/Desktop/Equating").  
# load data into R environment.  
Test1 <- read.table("test1.csv", sep = ",", header = FALSE).  
Test2 <- read.table("test2.csv", sep = ",", header = FALSE).
```

#### Session 2- Dimensionality assessment using DETECT index.

```
# call the package sirt.  
library(sirt).  
N <- 1210.  
I <- 50.  
b <- sample( seq( -2, 2, len = I)).  
Test1 <- sirt::sim.raschtype( stats::rnorm(N), b = b).  
Test2 <- sirt::sim.raschtype( stats::rnorm(N), b = b).  
# estimate Rasch model and corresponding WLEs.
```

```
mod1 <- TAM::tam.mml( Test1).
wmod1 <- TAM::tam.wle(mod1)$theta.
mod1 <- TAM::tam.mml( Test2).
wmod1 <- TAM::tam.wle(mod1)$theta.
# define item cluster.
itemcluster <- c( rep(1,50), rep(2,l-50)).
# compute Confirmatory DETECT statistic.
detect1 <- sirt::conf.detect( data = Test1, score = wmod1, itemcluster = itemcluster).
detect1 <- sirt::conf.detect( data = Test2, score = wmod1, itemcluster = itemcluster).
```

### Session 3 – Item calibration.

```
# from package mirt.
library(mirt).
# estimate a 3PL model for each group using the R package mirt.
m1 <- mirt(Test1,1,itemtype = "3PL").
m2 <- mirt(Test2,1,itemtype = "3PL").
# test calibration (item parameter).
PARA <- coef(m1, simplify = TRUE).
PARA1 <- coef(m2, simplify = TRUE).
print(PARA1).
```

### Session 4 – Equating of the two test forms.

```
# from package equateIRT.
library(equateIRT).
# import item parameters estimates and covariance matrices from mirt package.
estm1 <- import.mirt(m1,display = FALSE).
estm2 <- import.mirt(m2,display = FALSE).
estv <- list(estm1$var,estm2$var).
estc <- list(estm1$coef,estm2$coef).
# three-parameter logistic model.
test <- paste("test",1:2,sep = "").
mod3pl <- modIRT(coef = estc,var = estv,names = test,display = FALSE).
# direct equating coefficients between tests 1 and 2 using the Stocking-Lord method.
L12 <- direc(mods = mod3pl,which = c(1,2),method = "Stocking-Lord")
# extract equating coefficients of Stocking-Lord method.
summary(L12).
# scoring using direct equating coefficients and the observed score equating method.
SL <- score(L12,method = "OSE",se = FALSE).
print(SL).
# direct equating coefficients between tests 1 and 2 using the Haebara method.
H12 <- direc(mods = mod3pl,which = c(1,2),method = "Haebara")
# extract equating coefficients of Haebara method.
summary(H12).
# scoring using direct equating coefficients and the observed score equating method.
HH <- score(H12,method = "OSE",se = FALSE).
print(HH).
```

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Cook LL. Practical problems in equating test scores: a practitioner's perspective. In: Dorans NJ, Pommerich M, Holland PW, editors. *Linking and aligning scores and scales*. 2nd ed. New York: Springer; 2007. p. 73–87.
2. Tian F. A comparison of equating/linking using the Stocking-Lord method and concurrent calibration with mixed-format tests in the non-equivalent groups' common-item design under IRT. Published doctoral dissertation, Boston College. 2011. <https://core.ac.uk/reader/151481113>.
3. Clark LA, Watson D. Constructing validity: Basic issues in objective scale development. In: Kazdin AE, editor. *Methodological issues and strategies in clinical research*. Washington: American Psychological Association; 2016. p. 187–203.
4. Kline P. *A handbook of test construction (psychology revivals): introduction to psychometric design*. Oxfordshire: Routledge; 2015.
5. Atsua TG, Uzoeshi IV, Oludi P, Wagbara ES. Equating 2015 and 2016 basic education certificate examination on civic education using classical test theory and item response theory in Oyo State, Nigeria. *J Pristine*. 2018; 14(1).
6. Adewale JG. Equating two-year BECE results in Basic Science and Technology in Oyo State, Nigeria. 2015. Accessed December 14th 2022 from [www.buse.ac.zw/downloads/J.%20Gbenga%20Adewale.doc](http://www.buse.ac.zw/downloads/J.%20Gbenga%20Adewale.doc).
7. Fakayode O. Comparison of classical test theory and item response theory approaches in equating test scores from West Africa Examinations Council mathematics objective test. (Unpublished doctoral dissertation). University of Ibadan, Nigeria. 2018.
8. Babatimehin T. Determination of the equivalence of WAEC and NECO SSCE chemistry items using linear equating approaches of classical test theory and item response theory. *Bulgarian J Sci Educ Policy*. 2021;15(1):187.
9. Zannu BG. Equating 2012 and 2013 Physics test of unified tertiary matriculation examination of the Joint Admission and Matriculation Board. In: Adegoke BA, Popoola O, Babatunde OE, editors. *Public examining in subSaharan Africa: Issues, challenges and prospects*. Garki Abuja: Marveolus Mike Press Ltd; 2016. p. 197–212.
10. Kolen MJ. Comparison of traditional and item response theory methods for equating tests. *J Educ Meas*. 1981;18:1–11.
11. Schumacker RE. Test Equating. 2005. <http://www.appliedmeasurementassociates.com/white20papers/TESTEQUATING.Pdf>. Accessed 10 Nov 2022.
12. Ayanwale MA, Adeleke JO, Mamadelo TI. Invariance person estimate of basic education certificate examination: classical test theory and item response theory scoring perspective. *J Int Soc Teach Educ*. 2019;23(1):18–26.
13. Von Davier A. *Statistical models for test equating, scaling, and linking*. Berlin: Springer Science & Business Media; 2010. <https://doi.org/10.1007/978-0-387-98138-3>.
14. Chalmers RP. mirt: a multidimensional item response theory package for the R environment. *J Stat Softw*. 2012;48(6):1–29. <https://doi.org/10.18637/jss.v048.i06>.
15. Ayanwale MA. Efficacy of Item Response Theory in the Validation and Score Ranking of Dichotomous and Polytomous Response Mathematics Achievement Tests in Osun State, Nigeria. Ph.D thesis, University of Ibadan. 2019. <https://doi.org/10.13140/RG.2.17461.22247>.
16. Chong HY. A simple guide to the item Response theory (IRT) and Rasch Modelling. 2007. <http://www.creative-wisdom.com>. Accessed 6 Nov 2022.
17. Ayanwale MA, Adeleke JO. Efficacy of item response theory in the validation and score ranking of dichotomous response mathematics achievement test. *Bulgarian J Sci Educ Policy*. 2020;14(2):260–85.
18. Robitzsch A. sirt: Supplementary item response theory models. R package version 1.12.2. 2016.
19. Skaggs G, Lissitz R. An exploration of the robustness of four test equating models. *Appl Psychol Meas*. 1986;10(3):303. <https://doi.org/10.1177/014662168601000308>.
20. Kuhfeld M, Soland J. Avoiding bias from sum scores in growth estimates: an examination of IRT-based approaches to scoring longitudinal survey responses. *Psychol Methods*. 2022;27(2):234–60. <https://doi.org/10.1037/met0000367>.
21. Clifford S, Sheagley G, Piston S. Increasing precision without altering treatment effects: repeated measures designs in survey experiments. *Am Polit Sci Rev*. 2021;115(3):1048–65. <https://doi.org/10.1017/S0003055421000241>.
22. Meng Y. Comparison of kernel equating and item response theory equating methods (Unpublished doctoral dissertation). University of Massachusetts, Amherst. 2012.
23. Holland PW, Dorans NJ. *Linking and equating*. In: Brennan RL, editor. *Educational measurement*. 4th ed. Westport, CT: Praeger; 2006. p. 187–220.
24. Skaggs G. To match or not to match samples on ability for equating: a discussion of five articles. *Appl Measur Educ*. 1990;3(1):105–13. [https://doi.org/10.1207/s15324818ame0301\\_8](https://doi.org/10.1207/s15324818ame0301_8).
25. Woldbeck T. Basic concepts in modern methods of test equating. Paper presented at the annual meeting of the Southwest Psychological Association (New Orleans, April 11, 1998). Accessed October 23th 2022 from <https://eric.ed.gov/?id=ED417215>.
26. Kolen M, Brennan R. *Test equating, scaling, and linking: methods and practices*. 3rd ed. New York: Springer; 2014.
27. Ryan J, Brockmann F. A practitioner's introduction to equating with primers on classical test theory and item response theory. Council of Chief State School Officers. 2009. <https://files.eric.ed.gov/fulltext/ED544690.pdf>.
28. Kolen M, Brennan RL. *Test equating, scaling, and linking: methods and practices*. 2nd ed. New York: Springer; 2004.
29. Kolen MJ, Brennan RL. *Test equating: methods and practices*. Berlin: Springer Science & Business Media; 2013.
30. Dorans NJ, Pommerich M, Holland PW. *Linking and aligning scores and scales*. New York: Springer; 2007.
31. Von Davier A. *Statistical models for test equating, scaling, and linking*. New York: Springer; 2011. <https://doi.org/10.1007/978-0-387-98138-3>.
32. González J, Wiberg M. *Applying test equating methods using R*. New York: Springer; 2017. <https://doi.org/10.1007/978-3-319-51824-4>.
33. Van der Linden WJ. Equating error in observed-score equating. *Appl Psychol Meas*. 2006;30(5):355–78. <https://doi.org/10.1177/0146621606289948>.
34. Ragland S. An evaluation of a new method of IRT scaling. Harrisonburg: James Madison University; 2010.
35. Issayeva L, Thompson N. What is test equating? 2021. <https://assess.com/test-equating/>.



36. Dorans NJ, Moses TP, Eignor DR. Principles and practices of test score equating. *ETS Res Rep Ser.* 2010;2010(2):1–41.
37. Angoff WH. The College Board Admissions Testing Program: a technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Tests. 1971. Accessed January 10th 2023 from <https://eric.ed.gov/?id=ED050181>.
38. Liu M, Holland PW. Exploring population sensitivity of linking functions across three law school admission test administrations. *Appl Psychol Meas.* 2008;32:27–44. <https://doi.org/10.1177/0146621607311576>.
39. Chan RC, Shum D, Touloupoulou T, Chen EY. Assessment of executive functions: review of instruments and identification of critical issues. *Arch Clin Neuropsychol.* 2008;23(2):201–16. <https://doi.org/10.1016/j.acn.2007.08.010>.
40. Andersson B. Asymptotic standard errors of observed-score equating with polytomous IRT models. *J Educ Meas.* 2016;53(4):459–77.
41. Moses T. Linking and comparability across conditions of measurement: established frameworks and proposed updates. *J Educ Meas.* 2022. <https://doi.org/10.1111/jedm.12322>.
42. Altıntaş Ö, Wallin G. Equality of admission tests using kernel equating under the non-equivalent groups with covariates design. *Int J Assess Tools Educ.* 2021;8(4):729–43. <https://doi.org/10.21449/ijate.976660>.
43. Livingston SA. Equating test scores (without IRT). Princeton: Educational Testing Service; 2004.
44. Dong Y, Clements DH, Day-Hess CA, Sarama J, Dumas D. Measuring early childhood mathematical cognition: validating and equating two forms of the research-based early mathematics assessment. *J Psychoeduc Assess.* 2021;39(8):983–98. <https://doi.org/10.1177/07342829211037195>.
45. Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. *J Adv Nurs.* 2000;32(4):1008–15.
46. Setiawan R. A comparison of score equating conducted using haebara and stocking lord method for polytomous. *Eur J Educ Res.* 2019;8(4):1071–9. <https://doi.org/10.12973/eu-jer.8.4.1071>.
47. Van der Linden WJ. A test-theoretic approach to observed-score equating. *Psychometrika.* 2000;65(4):437–56. <https://doi.org/10.1007/BF02296337>.
48. Jang EE, Roussos L. An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *J Educ Meas.* 2007;44(1):1–21. <https://doi.org/10.1111/j.17453984.2007.00024.x>.
49. Kolen MJ. Linking assessments: concept and history. *Appl Psychol Meas.* 2004;28:219–26. <https://doi.org/10.1177/0146621604265030>.
50. Lissitz RW, Huynh H. Vertical equating for the Arkansas ACTAAP assessments: issues and solutions in determination of adequate yearly progress and school accountability. *Pract Assess Res Eval.* 2003. <https://doi.org/10.7275/npzw-wd59>.
51. Harris DJ. Practical issues in vertical scaling. In: Dorans NJ, Pommerich M, Holland PW, editors. *Linking and aligning scores and scales.* New York: Springer; 2007. p. 233–51.
52. Wu M. Measurement, sampling, and equating errors in large-scale assessments. *Educ Meas Issues Pract.* 2010;29(4):15–27. <https://doi.org/10.1111/j.1745-3992.2010.00190.x>.
53. Tong Y, Wu SS, Xu M. A comparison of pre-equating and post-equating using large-scale assessment data. In the annual meeting of the American Educational Research Association, New York, NY. 2008.
54. Lyrén P-E, Hambleton RK. Consequences of violated equating assumptions under the equivalent group's design. *Int J Test.* 2011;11(4):308–23. <https://doi.org/10.1080/15305058.2011.585535>.
55. Cook LL, Eignor DR, Taft HL. A comparative study of curriculum effects on the stability of IRT and conventional item parameter estimates (RR-85-38). Princeton NJ: Educational Testing Service; 1985.
56. Cook LL, Petersen NS. Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Appl Psychol Meas.* 1987;11:225–44. <https://doi.org/10.1177/014662168701100302>.
57. Petersen NS, Cook LL, Stocking ML. IRT versus conventional equating methods: a comparative study of scale stability. *J Educ Stat.* 1983;8:137–56. <https://doi.org/10.3102/10769986008002137>.
58. Budescu DV. Efficiency of linear equating as a function of the length of the anchor test. *J Educ Meas.* 1985;22:13–20.
59. Wingersky MS, Cook LL, Eignor DR. Specifying the characteristics of linking items used for item response theory item calibration. *ETS Res Rep.* 1987. <https://doi.org/10.1002/j.2330-8516.1987.tb00228.x>.
60. Hanson BA, Béguin AA. Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Appl Psychol Meas.* 2002;26(1):3–24. <https://doi.org/10.1177/0146621602026001001>.
61. Hambleton RK, Swaminathan H, Rogers HJ. *Fundamentals of item response theory.* USA: Sage; 1991.
62. Uysal I, Kilmen S. Comparison of item response theory test equating methods for mixed format tests. *Int Online J Educ Sci.* 2016;12(3):1–11. <https://doi.org/10.15345/ijoes.2016.02.001>.
63. Haebara T. Equating logistic ability scales by a weighted least squares method. *Jpn Psychol Res.* 1980;22(3):144–9.
64. Battauz M. IRT test equating in complex linkage plans. *Psychometrika.* 2013;78(3):464–80. <https://doi.org/10.1007/s11336-012-9316-y>.
65. Rahmawati R, Mardapi D. Modified Robust Z method for equating and detecting item parameter drift. *REID (Research and Evaluation in Education).* 2015;1(1):100–13. <https://doi.org/10.21831/reid.v1i1.4901>.
66. Kilmen S, Demirtasli N. Comparison of test equating methods based on item response theory according to the sample size and ability distribution. *Procedia Soc Behav Sci.* 2012;46:130–4. <https://doi.org/10.1016/j.sbspro.2012.05.081>.
67. Moghadamzadeh A, Salehi K, Khodaie E. A comparison method of equating classic and Item Response Theory (IRT): a case of Iranian study in the university entrance exam. *Procedia Soc Behav Sci.* 2011;29:1368–72. <https://doi.org/10.1016/j.sbspro.2011.11.375>.
68. Kim S, Lee WC. An extension of four IRT linking methods for mixed-format tests. *J Educ Meas.* 2006;43:53–76. <https://doi.org/10.1111/j.1745-3984.2006.00004.x>.
69. Battauz M. Testing for differences in chain equating. *Statistica Neerlandica.* 2022. <https://doi.org/10.1111/stan.12277>.
70. Oguoma CC. Assessment of the equivalence of WAEC and NECO mathematics multiple-choice tests using item response theory. 2017;1–231. Accessed 10th January from [https://phd-dissertations.unizik.edu.ng/repos/81101585000\\_155715043200.pdf](https://phd-dissertations.unizik.edu.ng/repos/81101585000_155715043200.pdf).
71. Adeyemo EO. Comparability of item parameters of equivalent examinations using linear test score equating method. *Int J Educ Res.* 2019;7(2):255–66.
72. Amusa JO, Ayanwale MA, Oladejo AI, Ayedun F. Undergraduate physics test dimensionality and conditional independence: perspective from latent traits model package of R language. *Int J Assess Eval.* 2022;29(2):47. <https://doi.org/10.18848/2327-7920/CGP/v29i02/47-61>.
73. Kivunja C, Kuyini AB. Understanding and applying research paradigms in educational contexts. *Int J High Educ.* 2017;6(5):26–41.

74. National Business and Technical Examinations Board (NABTEB). The Guardian Nigeria. 11 May 2021. Accessed 23 Sept 2022.
75. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. 2016.
76. Robitzsch A, Robitzsch MA. Package 'sirt'. 2022. <https://cran.r-project.org/web/packages/sirt/sirt.pdf>.
77. Stout W, Habing B, Douglas J, Kim HR. Conditional covariance-based nonparametric multidimensionality assessment. *Appl Psychol Meas*. 1996;20(4):331–54. <https://doi.org/10.1177/014662169602000403>.
78. Zhang J. A procedure for dimensionality analyses of response data from various test designs. *Psychometrika*. 2013;78(1):37–58. <https://doi.org/10.1007/s113360129287z>.
79. Ayanwale MA, Isaac-Oloniyo F, Abayomi FR. Dimensionality assessment of binary response test items: a non-parametric approach of Bayesian item response theory measurement. *Int J Eval Res Educ*. 2020;9(2):385–93. <https://doi.org/10.11591/ijere.v9i2.20454>.
80. Zhang J, Stout W. Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika*. 1999;64(2):129–52. <https://doi.org/10.1007/BF02294532>.
81. Zhang J, Stout W. The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*. 1999;64(2):213–49. <https://doi.org/10.1007/BF02294536>.
82. Zhang J. Conditional covariance theory and DETECT for polytomous items. *Psychometrika*. 2007;72(1):69–91. <https://doi.org/10.1007/s1133600412577>.
83. Bonifay WE, Reise SP, Scheines R, Meijer RR. When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. *Struct Equ Model*. 2015;22(4):504–16. <https://doi.org/10.1080/10705511.2014.938596>.
84. Battauz M. equateIRT: Direct, Chain and Average Equating Coefficients with Standard Errors Using IRT methods. R package version 1.2-2, URL. 2015. Accessed 27th February from <http://CRAN.R-project.org/package=equateIRT>.
85. Lord F. Applications of item response theory to practical testing problems. Hillsdale: Lawrence Erlbaum Associates; 1980.
86. Hambleton RK, Swaminathan H. Item response theory: principles and applications. Dordrecht: Kluwer Nijhoff Publishing; 1985.
87. De Boeck P, Wilson M. Explanatory item response models: a generalized linear and non-linear approach. New York: Springer; 2004.
88. Van der Linden WJ, editor. Handbook of item response theory. Three volume set. Boca Raton: Chapman and Hall/CRC; 2016.
89. Cook LL, Eignor D. IRT equating methods. *Educ Meas Issues Pract*. 1991;10(3):37–45. <https://doi.org/10.1111/j.1745-3992.1991.tb00207.x>.
90. Tobih DO, Ayanwale MA, Ajayi OA, Bolaji MV. The use of measurement frameworks to explore the qualities of test items. *Int J Eval Res Educ*. 2023;12(2):914–23. <https://doi.org/10.11591/ijere.v12i2.23747>.
91. Dimitrov DM. An approach to scoring and equating tests with binary items: piloting with large-scale assessments. *Educ Psychol Measur*. 2016;76(6):954–75. <https://doi.org/10.1177/0013164416631100>.
92. Ayanwale MA, Chere-Masopha J, Morena MC. The classical test or item response measurement theory: the status of the framework at the examination Council of Lesotho. *Int J Learn Teach Educ Res*. 2022;21(8):384–406.
93. Stocking ML, Lord FM. Developing a common metric in item response theory. *Appl Psychol Meas*. 1983;7(2):201–10. <https://doi.org/10.1177/014662168300700208>.
94. De Mars C. Item response theory. Understanding statistics measurement. Oxford: Oxford University; 2010.
95. Baker FB. The basics of item response theory. Test Calibration. ERIC Clearinghouse on Assessment and Evaluation. Accessed 12th February 2023 from <https://eric.ed.gov/?id=ED458219>. 2001.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.