International Journal of Learning, Teaching and Educational Research
Vol. 21, No. 8, pp. 384-406, August 2022
https://doi.org/10.26803/ijlter.21.8.22
Received Apr 16, 2022; Revised Jul 29, 2022; Accepted Sep 1, 2022

The Classical Test or Item Response Measurement Theory: The Status of the Framework at the Examination Council of Lesotho

Musa Adekunle Ayanwale*, Julia Chere-Masopha and Malebohang C. Morena
National University of Lesotho, Roma, Maseru, Lesotho

Abstract. While the Examination Council of Lesotho (ECOL) is burdened with a huge workload of assessment tasks, their procedures for developing tests, analysing items, and compiling scores heavily rely on the classical test theory (CTT) measurement framework. The CTT has been criticised for its flaws, including being test-oriented, sample dependent, and assuming linear relationships between latent variables and observed scores. This article presents an overview of CTT and item response theory (IRT) and how they were applied to standard assessment questions in the ECOL. These theories have addressed measurement issues associated with commonly used assessments, such as multiple-choice, short response, and constructed response tests. Based on three search facets (Item response theory, classical test theory, and examination council of Lesotho), a comprehensive search was conducted across multiple databases (such as Google Scholar, Scopus, Web of Science, and PubMed). The paper was theoretically developed using the electronic databases, keywords, and references identified in the articles. Furthermore, the authors ensure that the keywords are used to identify relevant documents in a wide variety of sources. A general remark was made on the effective application of each model in practice with respect to test development and psychometric activities. In conclusion, the study recommends that ECOL switch from CTT to modern test theory for test development and item analysis, which offers multiple benefits.

Keywords: classical test theory; item response theory; Examination Council of Lesotho; item development; item analysis

1. Introduction

The Examinations Council of Lesotho (ECOL), the central body for all examinations and assessments in Lesotho, is located right in the heart of Maseru,

* Corresponding author: *Musa Adekunle, Ayanwale; ma.ayanwale@nul.ls*

the capital. The Examinations Council Regulations were enacted in 1986. It was then configured as a unit of the Ministry of Education and Training (MOET), responsible for conducting public examinations at the national level. ECOL is a non-profit organisation that undertakes various functions, including the control and arrangement of public examinations, the issuance of certificates to all successful applicants, and any other things necessary or incidental to the proper administration and functioning of the Council (Exam Council of Lesotho, 2018). In 2003, the ECOL's mandate was further expanded to include assessing the educational system's performance and developing continuous assessments for usage by providing school materials to facilitate the assessments. Therefore, seeking to become the world's premier assessment institute offering high-quality services, maintaining high standards of quality in professional education, and ensuring that integrity is maintained in executing their core responsibility of internationally recognised certifications to students at the pre and basic levels. However, ECOL's responsibilities encompass a range of activities, including the development, implementation, monitoring, and evaluation of an appropriate, fair, and reliable education assessment system for elementary and secondary schools, as well as awarding internationally credentialed qualifications that become part of the education system (Exam Council of Lesotho, 2018).

ECOL also administers level evaluations at specific points in the education system, such as in Grade 7 when students take their Primary School Leaving Examination (PSLE), in Junior Secondary, Form C (Junior Certificate Examination), and in Senior Secondary, Form E (O'Level/LGCSE). Furthermore, it conducts the National Assessment Survey in partnership with the National Curriculum Development Center at the elementary level (Grades 4 and 6). Every two years, these surveys are conducted to evaluate the educational system's performance throughout the country in numeracy and literacy. It is worth noting that the Cambridge International Assessment accredits the O' Level curriculum and examinations. As well as administering exams, the Council acts as an agent for other international testing bodies, such as the University of London and the Management College of Southern Africa (MANCOSA). Despite the overwhelming assessment tasks that ECOL is saddled with, experience and mode of operation show that their procedures for test development, item analysis, and scoring framework are heavily reliant on the classical test theory (CTT) method of measurement, which has been criticised for its shortcomings, such as test-oriented rather than item-oriented, assumes linear relations between latent variables and observed scores, hence it is impossible to estimate the true score directly, or without making strong assumptions, item parameters such as discrimination and difficulty of the test items depend on the sample used, and the standard error of measurement, a function of test score reliability and variance, is universal for all examinees. These limitations can pose several challenges when used in high-stakes exams such as ECOL. For instance, CTT fails to account for observed distributions of test scores that have the floor or ceiling effects, in which a large proportion of examinees score near the low or high end of the range (Demars, 2017; Jabrayilov et al., 2016; Rusch et al., 2017). Due to difficulties in resolving these problems within the framework of classical measurement theory, the measurement community and assessment

organisations have switched to a modern theory known as item response theory (IRT) (Embretson & Reise, 2013).

CTT's shortcomings were addressed with the development of modern theory, which allows for non-linear relationships, estimation of the true score independent of the sample used, sample invariant estimation of parameter values, and gives an assessment expert the ability to select items that are in accordance with the desired model and applies internal consistency and reliability concepts to derive more information about how measurements are conducted. IRT has also been established as an essential tool for test development, item analysis, and evaluation, which leads to precise, valid, and relatively less burdensome instrument responses (Edelen & Reeve, 2007). The plethora of studies have confirmed that the IRT framework offers a multitude of advantages that have sparked the interest of educational assessment institutions, test developers, and policymakers in the assessment industry, who have adopted it for valid and reliable decision making (Ayanwale et al., 2019; Cai et al., 2016; Embretson & Reise, 2013; Ewing et al., 2005; Ganglmair & Lawson, 2010; Hambleton & Swaminathan, 1985; Lang & Tay, 2021). To date, ECOL has yet to embrace and integrate the potential of IRT in educational assessments and testing despite its promising development. The purpose of this paper is to fill this gap by encouraging ECOL to incorporate IRT into their existing methodologies by providing an overview of each measurement theory, its assumptions, its models, weaknesses, and strengths to improve the assessment and scoring procedures currently used by ECOL, which in turn enhances the validity of the certificate awarded.

2. Classical Test Theory

Imagine examinees are given 20 questions. Sixteen of the 20 questions are equally hard; two are difficult while two are easy. The two examinees get 18 test items correctly. Both get 90%. Examinee "A" has made two simple mistakes, while examinee "B" has made two very complicated ones. How can we determine which examinee has more ability? This scenario highlights a significant flaw in the CTT method of testing. Historically, CTT refers to a theory of test scores in which three elements (observe, true, and error scores) are introduced (Hambleton & Jones, 1993; Steyer, 2001). Models of various forms have been developed within the theoretical framework.

In the classical test model, two unobservable variables are linked to an observable test score (X), true score (T), and error score (E), that is: $X = T + E$. True score cannot be observed directly; It can only be estimated from an examinee's responses to a set of items whose responses correspond to the actual abilities that particular examinees possess, though there are inherent errors in estimation. Factors such as fatigue, guessing, or stress can cause random errors (Bovaird & Embretson, 2012). Examinees' observed scores represent their total scores on a test. It would have been the true score if not for the error score. Standard error of measurement (SEM) plays a major role in CTT, which are standard deviations of measurement errors for each group of examinees. A test's variability or spread can be determined from its measurement errors. In $X = T + E$, the true score equals the average of a person's observed scores and accounts

for measurement error. Because measurement error cannot be determined, every standardised test has an SEM. SEM is measured in standard deviations. In this way, the reliability of the test is determined. Precision and reliability of measurements are higher with a smaller SEM. The error in CTT is conceived as random and non-systematic. Several factors, internal or external to the examinee, may account for it. Test items created poorly or tested under poor conditions are examples of external errors. Internal errors are those caused by the examinee, such as fatigue, stress, and a lack of concentration (Ayanwale, 2019).

Item and test level statistics are part of CTT. Item difficulty and discrimination are analysed at the item level. Item difficulty index is represented by 'p' and indicates the proportion of correct answers. The item discrimination index is indicated by a 'D'; it tells us how distinct the item is between those with high and low abilities. CTT looks at the reliability of parallel tests (Demars, 2017). A parallel test measures the same latent ability with the examinees having the same true score and errors on both tests. Many items are generated that represent a single content domain for parallel tests. Ideally, this set should have twice the number of items intended for a single test form (Brown, 2013).

2.1 Assumptions of classical test theory

In CTT, three assumptions are made. First, the correlation between the error and true scores is zero. In this case, the variance of a true and error score is equal to the variance of the observed score, which is true if $\gamma Te = 0$ (Steyer, 2001). In the equation Var(X) = Var(T) + Var(E), Var(.) is the variance, while the reliability Rel(X) is defined as

$$\text{Rel(X)} = \frac{Var(T)}{Var(X)} = \frac{Var(T)}{Var(T)+Var(E)} \qquad \text{----------Eqn. 1}$$

Consequently, correlation coefficients between two parallel measurements determine the reliability of the CTT test. Adedoyin (2010) argues that error variance decreases as measurement reliability increases. When the error variance is small, the observed score of test-takers is close to the true score. However, when error variance is large, observed scores do not always reflect true scores (IResearchNet, 2022). The second assumption says errors have a zero mean. Thus, these random errors are expected to cancel out over many repeated measurements, resulting in a zero expected mean error rate. The observed score equals the true score once an error is zero, (X=T), $\sum_{i}^{n} \frac{\mathrm{E}}{\mathrm{N}} = 0$

A third assumption is that measurements from parallel are uncorrelated. A parallel test is defined in classical test theory as two measures of X and X^1 that have the same true score ($T=T^1$) and the same observed variances $\delta^2(X) = \delta^2(X^1)$. Ojerinde (2013) suggests that two tests can be considered parallel if the expected values of X and X^1 are equal (that is, $E(X) = E(X^1)$). There is typically an equal error variance for the two parallel scores if $X \parallel X^1$ if $X_1 = X_2 = T_i + E_i$ for every population of tests.

2.2 Classical test theory method of item analysis

Test items are analysed quantitatively and qualitatively to determine their characteristics. To facilitate instrument improvement, the purpose is to revise or discard items that do not meet minimally acceptable standards. In item development, it is crucial to consult experts who possess a mastery of relevant materials. Experts and review boards find it difficult to determine the quality of "poor" items because of the test content's multidisciplinary nature and examinees' demographics (Krishnan, 2013). Data analysis helps identify issues that slipped experts' attention. The goal of item analysis is to select items that maximise reliability. Matching what is taught with what is assessed is crucial. There should be a mixture of basic and advanced knowledge in any exam. Examinees become frustrated if items are too difficult, while overconfidence and a decline in motivation are consequences of too easy assessments (Esmaeeli et al., 2021). Creating item banks that are reusable is important through item improvement. Ayanwale et al. (2019); Crocker and Algina (1986) defined item analysis as evaluating test items for test construction and revision. This is a technique for improving test items. In addition to identifying biased or unfair items, item analysis can also identify poorly worded questions (Grand et al., 2013; Khan et al., 2013). Results of item analysis are then used to refine the items of interest. Revision is needed for items that are more difficult or too easy. In addition, test scores can be observed to enhance item analysis by observing their reliability, although the literature on measurement discusses item analysis separately from reliability. To establish test scores' reliability, item difficulty and discrimination are essential components of item analysis(Elgadal & Mariod, 2021; Toksöz & Ertunç, 2017).

2.3 Parameter estimation of Classical test theory

Item difficulty is an important concept in CTT. For DeVellis (2006), it is the percentage of examinees who answered an item correctly. In CTT, item difficulty is sample-based. These values are invariant only for groups of similar level examinees. CTT often refers to item difficulty as a p-value. Divide the number of respondents who selected a particular answer by the total number of respondents in the sample to find the percentage of those deciding to pick that response, and you get a p-value for each response and the correct answer. The p-values can be expressed mathematically as:

$$p = \frac{\text{number of an examinee who got the item right}}{\text{total number of an examinee who attempted the items}}$$

The proportion of examinees that got the item wrong can be expressed as:

$$q = \frac{\text{number of an examinee who got the item wrong}}{\text{total number of an examinee who attempted the items}}$$

Hence, pq is the variance, and ($SD = \sqrt{pq}$) is the standard deviation. The item difficulty index (p) ranges between 0 and $p \leq 1$. A value of 1 is considered to be very simple if all members of the sample correctly answered the question, while a p-value of 0 is indicative of none of the respondents in the sample answering the question correctly; such an item is said to be hard (Cappelleri et al., 2014; Kline, 2014). For Courville (2005), Items with dichotomously scored items have a

greater item variance (that is, $\sigma_i^2 = p_i q_i$), indicating the importance of the item difficulty (p) in the variance measure, while (q) indicates the significance of the item type (difficulty). The items variance and the total variance of the result are thus representations of item difficulty.

Further, Crocker and Algina (1986) pointed out that the item difficulty of a norm-referenced test usually falls between 0.60 and 0.80. The reason is due to the item format typically used on such tests. Open-ended questions have a remote or zero likelihood of being answered correctly. The probability of guessing correctly increases when the test format is multiple choice. As a result, p is the proportion of respondents who know the answer (p), and 1/m is the number of responses that reflect how many of those who didn't know the answer but correctly guessed (m) responded. As a multiple-choice test, we do not aim to maximise item difficulty at p=0.50; instead, we aim to maximise item score variability. Hence, item difficulty should optimise item score variability. The proportion of correct answers is 1/m, which is known. In addition, item variation at 0.50 is the optimal level; the p-values of items with maximum true score variance also vary due to examinees' random guessing. This can be written as $p^1 = 0.5 + \frac{0.5}{m}$, where p^1 is the observed p-value, and m is the number of alternatives or distracters.

However, the item difficulty index that maximises item variance is (p^1=0.5+0.125) = 0.63 for multiple-choice items with four (4) options, and (p^1=0.5+0.1) = 0.60 for five (5) options (Cohen & Swerdlik, 2009; Cohen et al., 2013; Filgueiras et al., 2014; Hill et al., 2013). Items in a test with a difficulty level higher or lower than 0.60 and whose difficulty level exceed or fall below 0.63 should be deemed inappropriate. In traditional norm-referenced testing, items with a difficulty index greater than 0.70 or less than 0.30 are considered bad items (Adegoke, 2013; Hambleton & Jones, 1993).

Item discrimination is another CTT parameter. It indicates that the examinee's ability differs. Generally, high, average, and low scores are expected. Among the purposes of analyzing test items is selecting items that can separate examinees into different categories with respect to their abilities. High-ability examinees should be able to score a test item correctly, while low-ability examinees will score it incorrectly. Test items that have such properties are discriminatory by nature. Criterion scores place examinees in upper or lower groups based on their total test scores. This grouping of examinees makes the discrimination index controversial (Algina & Swaminathan, 2015; Rusch et al., 2017). The lower group had 50% participants, while the higher group had 50%. A criterion of interest is easily distinguished between very high and very low scores.

For Kelley (1939), cited in Ayanwale (2019), suggested that instead of 50%-50%, the item discrimination statistic would function correctly with a 27%-27% split since it would omit 46% of the data. As the sample size increases, the same statistic becomes as stable and useful when using a 27%-27% split (Crocker & Algina, 1986). A high score on a particular item usually indicates an examinee who has done well on the test. Hingorjo and Jaleel (2012); Vyas and Supe (2008)

suggest that items with negative discrimination should be revised or discarded if they are selected by a larger percentage of the lower scoring group than the higher scoring group. If an item is high- or low-performing, item discrimination can be calculated as $D = p_u - p_l$, where p_u is the proportion of correct answers for the upper group, and p_l is the proportion of correct answers for the lower group. After identifying the top 27% and bottom 27% of examinees, the percentage passing for each item is calculated for each group. The item discrimination index is obtained by subtracting the 'p' of the lower-performing group from the top-performing group. The index ranges from 0 to 1. A classic interpretation of item discrimination is provided by (Ebel, 1965):
1. If $D \geq 0.40$, very well-functioning item.

2. If $0.30 \leq D \leq 0.39$, reasonably well-functioning item.

3. If $0.20 \leq D \leq 0.29$, marginal items need to be revised.

4. If $D \leq 0.19$, a poorly functioning item needs to be expunged or fully revised.

More importantly, a discrimination index provides information about how an item differs on a certain criterion. This is problematic since it ignores a lot of data. For example, several examinees are omitted (46% of respondents), and information regarding examinees in the higher and lower groups (Courville, 2005). The product-moment correlation coefficient is applicable when the total and item scores are interval scales. A point-biserial correlation between dichotomous scored items and the total score is employed to resolve the problem (Adegoke, 2013). It measures the direction of the linear relationship of one factor with another that is continuous (Privitera, 2012). In point-biserial notation, $p_{bis} = \left[\frac{(\mu_y - \mu_x)}{\sigma_x}\right]\sqrt{p/q}$ with μ_y is the criterion score mean for the proportion of respondents answering the question correctly, while μ_x is the overall criterion score. A correlation coefficient between an item's performance and an examination's performance is also used to establish item discrimination (Brown, 2013; DeVellis, 2006). As a result, p-bis represents the correlation between items and total scores. The correlation should be positive since it demonstrates that correct answer holders scored higher and incorrect answer holders scored lower. If negative, you should revise or discard the items. The higher the value, the stronger the discrimination.

2.4 Reliability of scores in the context of classical test theory
The reliability of a test is the ability for identical scores to be achieved over a specified period whenever the same population of test subjects is examined (Demars, 2017). A reliability coefficient is expressed numerically, and any value around 0.70 and above is a good estimation of the reliability coefficient for an instrument (Preston et al., 2020). Tests with perfect reliability are seldom available, that is, tests capable of reproducing the same scores when administered to the same group. The observed results of a highly reliable test are close to its true scores. Therefore, using the square of the correlation between the observed and true score, the reliability coefficient can be calculated (Birnbaum,

1968). For Dent et al. (2001), true score variance is considered when determining reliability. Reliability estimates are based on random measurement errors and can be categorised into different types (Gay et al., 2011).

Using the test-retest method of reliability has two shortcomings. It is costly and time-consuming to administer the instrument for the first time, let alone for a second time. The resulting higher cost is far more concerning. If the sample population is high in mortality, it is more difficult to assess the reliability (Crocker & Algina, 1986). Secondly, the test-retest method can cause reactivity, as described by (Downing, 2003). Reactivity occurs when repeated testing eventually leads to a substantive change. In testing, memory is the main cause of reactivity. The memory may impact performance on the second test from the first test. Alternative reliability tests have been developed to solve these issues. A correlation is established between two similar tests administered to the same group (Crocker & Algina, 1986).

The test-retest method has reactivity problems which the alternative form method solves but has its problems. This method has a significant flaw because it is impossible to guarantee that each test samples the same content. This happens whenever you use two tests. To solve this problem, a single test reliability coefficient was developed. One administration of a single test is a method of estimating reliability. As a method of assessing reliability, internal consistency relies on the extent to which items within a single test are consistent with each other and the test overall. Split-half reliability is appropriate for long or hard tests, and Kuder-Richardson reliability (KR-20) is only appropriate for items with dichotomous scores, like selection-response tests. By using the split-half method, a test is given to all samples at once, then the test is divided into two parts, and the parts are compared (Crocker & Algina, 1986; Jabrayilov et al., 2016) claimed that splitting the test in many ways won't produce a unique estimate of reliability. This caused an important issue in reliability. Spearman (1910) developed the Spearman-Brown formula to estimate the reliability coefficient for the scores on the whole test to correct the pitfalls associated with split-half correlation.

For reliability estimation, item covariance methods are the most commonly used. The Cronbach alpha coefficient is the main method used to measure the internal consistency of a test or scale in the psychology and education fields (Demars, 2017). Alpha is merely a measure of precision and is not a measure of stability (Crocker & Algina, 1986). Kuder Richardson 20 (KR20) is the second item covariance analysis. Each item in the test is rated between 0 and 1. This score indicates how items in a given test measure the same construct or concept—the alpha coefficient increases when test items are highly correlated. Testing reliability and alpha are not only affected by correlation, but also depends on the length of the test. Therefore, a low value of alpha may reflect poor inter-item correlation or a long test. Mona (2014) recommends eliminating items with poor correlation or revising them.

2.5 Merits of classical test theory

CTT remains popular among educators despite new approaches to measuring proficiency (De Champlain, 2010). Its basic concepts are straightforward. Among its advantages, it makes relatively weak assumptions. The assumptions in CTT enable it to be applied to a wide variety of data. Anyone with basic mathematics skills can quickly grasp the concepts, as they aren't mathematically demanding. Cronbach's alpha measures reliability. CTT can be used to conduct the analyses with the common statistical packages. Psychometricians in education and psychology find it more acceptable.

Further, CTT-based measurements of instruments easily fit into underlying models, thus yielding desirable results. CTT is appealing because individual items don't have to be optimal, even if they relate only partially to an underlying construct; the concern can be alleviated by creating several items that assess the construct. Researchers have found that reliability can be improved to any desired level by increasing the number of items on a specific test concerning a variable (Wells & Wollack, 2018).

2.6 Limitations of classical test theory

For Rusch et al. (2017) noted that the assessed sample of examinees influences both item difficulty and discrimination indices. The study of Kolen (1981) found that the difficulty index is higher for examinees with high ability. In CTT, item difficulty has a bearing on examinee ability scores. Observed test scores are higher if the items are easy and lower if they are difficult. Another flaw in CTT assumes that all examinees have the same measurement error. A test's type influences test scores and true scores. The items on the test determine what students' scores will be. It is still possible to score lower on difficult tests and higher on easier ones, even though one has the same ability. Depending on each student's ability level, scores differ in error amounts. CTT also has the limitation that the same items must be used to compare examinees' performance. Parallel forms are difficult to achieve in CTT, further aggravating this limitation. Test reliability depends on parallel testing, which is based on a sample provided by the examinee.

For Traub (2015) argues reliability is a useful index of a test score's quality. Such an indicator depends on the characteristics of the group of test-takers. It is also test-oriented, making it difficult to predict examinees' responses to a test item (Crocker & Algina, 1986). Test developers cannot predict a test taker's performance on a particular item based on the CTT model. Examinee and item dependence is the most significant limitation of CTT. They are both affected by changes in the other's characteristics. Hence, comparing the characteristics of different tests and items taken by different groups of students is difficult.

3. Paradigm shift from classical test to item response theory

Several new measurement methods are being developed due to the limitations discussed above. In CTT, the group dependence, item-examinee ability mismatch, weak assumptions, and parallel testing problems present limitations. As an alternative, item response theory (IRT) or latent trait theory provides a solution to CTT's shortcomings (Bovaird & Embretson, 2012). Many other

models focusing on measurement issues developed an alternative model. Since IRT focuses on the item, all statistical analyses are done at the item level. It is one of the greatest advantages over CTT. Numerous studies in the fields of education and psychology have highlighted the same concept (Cappelleri et al., 2014; Embretson & Reise, 2013; Tay et al., 2015). The evidence demonstrates that IRT is widely used in these fields, and medical education is no exception (De Champlain, 2010; Downing, 2003; Preston et al., 2020).

Moreover, IRT is widely used to develop valid and accurate data about students' learning competencies in testing centers worldwide. The CTT assumptions were challenging to test and apply to practical problems, leading to alternative measurement models. The models are essentially extensions and liberalisations of conventional test theory. In addition, IRT is a necessary tool that has to be available in any large-scale testing center that requires a valid and reliable instrument.

4. Item response theory

IRT is a statistical model that describes both examinee items and test performance and further explains how the test results relate to the abilities reflected in the items on the test (Embretson & Reise, 2013). Responses to items may be discrete, continuous, or dichotomous. A score category may be ranked or unranked. One or more abilities may explain test scores. A variety of models can explore the relationship between item responses and underlying abilities. IRT has established and applied several models to test data. Hambleton and Swaminathan (1985) analysed the characteristics of item models as follows: the model should describe how observed responses relate to underlying non-observable constructs, the model should provide an estimation of the underlying construct, the examinee's scores must help estimate the construct, and the performance of an examinee must be completed predicted or explained by the underlying constructs. According to item response theory, an examinee has some unobservable, latent abilities that cannot be studied directly. IRT is used to develop models to relate latent traits to observable characteristics of an individual, especially their abilities to correctly answer questions in a test (Baker & Kim, 2017; Magis, 2007).

IRT employs mathematical functions, unlike classical test theory (CTT), which uses the model X=T+E. Based on Hambleton and Swaminathan (1985), IRTs are characterised by a strict relationship between responses and traits. Further, IRT is based on the assumption that one or more examinees' abilities can be predicted from theta (θ), which constitutes one of the parameters. Additionally, Crocker and Algina (1986) found that the observed score and the ability parameter are related to the observed and true scores. Their study highlighted that item difficulty and discrimination do not depend on examinee characteristics. Additionally, the ability estimates are likewise independent of the items and can be described as item-free, while the ability parameters are person-free.

4.1 Assumptions of item response theory

IRT models are fitted to data based on assumptions about the data (Edelen & Reeve, 2007). Assumptions in test theory state that an examinee will answer the question correctly if they know the correct answer. Without this assumption, testing is not justified. Other assumptions include local independence, dimensionality, and monotonicity. The assumptions should hold regardless of the latent trait method employed. These assumptions must be met for a test set to be meaningful when estimating latent trait models (Bichi & Talib, 2018; Zhu & Lu, 2017).

4.1.1. Dimensionality

A set of latent traits can explain test performance. A vector of ability scores can describe an examinee's ability in n-dimensional (θ_1, θ_2, θ_3….. θ_n). Items responding to the test with more than one latent trait are called multidimensional items, while items responding to the test that assumes one latent trait are known as unidimensional items. Only one area of knowledge, ability, or construct is measured in the items (Tay et al., 2015). The items on a one-dimensional test reflect only one dimension. One-score tests implicitly imply that the items share an overarching primary construct. In this model, each examinee is assigned single theta *, and uncontrolled variables can affect item responses as nuisance dimensions unique to the item and not shared by other items (Adewale et al., 2017). A test or ability scale containing all its items must measure a single latent attribute of an individual. Violating this assumption may lead to misleading results (Immekus et al., 2019).

In their study, Ojerinde and Ifewulu (2012) identified multiple methods for testing unidimensionality, such as the Cronbach analysis test, exploratory factor analysis, eigenvalue test, random baseline test, biserial test, factor loading test, congruence test, congruency or part-to-whole test, and vector frequency test, as well as confirmatory factor analysis. Various methods exist for assessing the unidimensionality of test data, depending on the nature of the test data. Predictive continuous and normally distributed data are tested for unidimensionality through parallel analysis, which VistaParan and MPLUS implement, or confirmatory factor analysis based on Pearson's correlation matrix (Adewale et al., 2017; Kline, 2005) implemented in AMOS or LISREL. Generally, polychoric correlation can be used parallel (implemented in FACTOR; Vista-Paran) when the data is ordinal (Metibemu, 2017). In dichotomously scored data, nonlinear factor analysis implemented in normal Ogive harmonic robust moment (NOHARM), parallel analysis based on tetrachoric correlation matrix (implemented in Vista-Paran), full information item factor analysis (implemented in EQSIRT, MIRT R package, and TESTFACT), bootstrap modified parallel analysis test (implemented in Itm R package), and stout essential dimensionality test (implemented in DIMPACK package) can be used (Ackerman, 2010; Finch & Monahan, 2008; Finch & French, 2015; Reckase, 2009). The next IRT assumption is local item independence.

4.1.2 Local item independence

Local item independence means that the chance of an examinee getting an item right is not affected by how they answered other items on the test. The fact that

students perform independently on different items does not mean they do not correlate; their abilities determine their performance. An examinee's probabilities are associated with a set of items related to the probability of a response pattern on that set of items. An ability is constant at a particular measurement time when it influences responses to a set of items. Therefore, the relationship between the two items should be as close as possible to zero. The responses may, therefore, be influenced by factors other than what the instrument was designed to measure. Given an individual's score on the latent trait, the observed items should be independent of each other (Debelak & Koller, 2020; Song et al., 2019). Independent means are statistically independent. Statistically, independent items exhibit their qualities and consider examinees' abilities to unfold their characteristic functions about them (Behavior et al., 2012; Yen, 2006).

Several approaches assess whether local item independence is valid (Debelak & Koller, 2020; Kim et al., 2011). These methods include the likelihood ratio G^2, the power-divergence (PD) statistic, the Q_3 statistic, Fisher's r-to-z transformed Q_3, the Wald test, the likelihood ratio test in logistic regression (LR G^2), the absolute value of mutual information difference (Tsai & Hsu, 2005), the mutual information difference (MID), the modification index (MI) in structural equation modelling (SEM), and the use of the residual correlation from the factor analysis (FA). Among the methods, only the likelihood ratio G^2 method is implemented in a popular IRT computer program such as item response theory-Patience response outcomes (IRTPRO). For Chen and Thissen (1997); Tang et al. (2020) proposed that the local dependency (LD) χ^2 statistic be computed by comparing the observed and expected frequencies in each of the two-way cross-tabulations between response to each item and each of the other items. Standardised χ^2 values (roughly Z-scores) become large when a pair of items indicate local dependency (Chen & Thissen, 1997). Additionally, an LD number greater than 10 signals local dependence (Adewale et al., 2017; Gay et al., 2011). The study (Yen, 1993) suggested Yen's Q-3 statistic as an effective measure for assessing local independence; after controlling for person location estimates, the Q3 statistic is the correlation of residuals between two items. The next IRT assumption is monotonicity.

4.1.3 Monotonicity

A normal ogive is the item response function (IRF). Item response curves have a mean of 0 and a standard deviation of 1. Item response functions are also known as item characteristics curves. Items characteristic curves (ICC) relate the probability of success on items to the ability measured by the item. In Birnbaum (1968); Lord (2012), ICC is invariant across groups of test takers, resulting in the invariance of item parameters that produce the item characteristic curve. This aspect is a prominent distinguishing feature of IRT compared to CTT.

The study of Hambleton and Swaminathan (1985) argues that invariance of item characteristics and ability parameters means that characteristics of an item do not depend on the abilities of examinees, just as characteristics of examinees do not depend on test items. ICC represents non-linear regressions between item score and latent trait. Because the variable and probability are unbounded, the

relationship will be nonlinear. It shows the probability of answering a question correctly as a function of ability. No matter the distribution of examinees, the probability is constant. In this case, the ICC will take the shape of a normal ogive since the probability remains the same no matter how many other examinees are nearby. The ICC has three sections: the lower asymptote, the upper asymptote, and the middle part. An ICC might require several parameters depending on the logistic model, as shown in Figure 1.

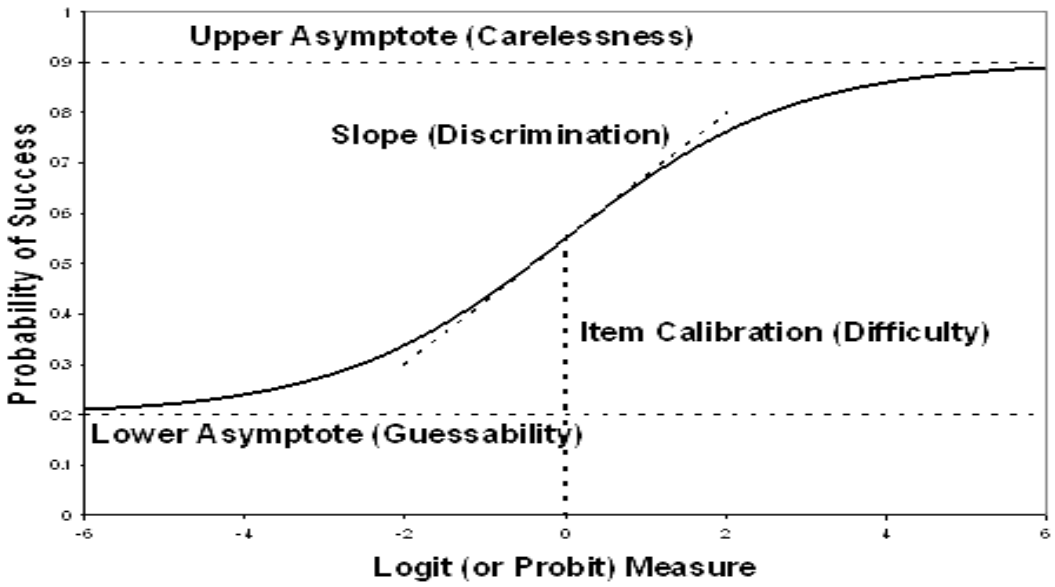

Figure 1: ICC showing parameters

Source: (Ojerinde & Ifewulu, 2012)

ICC curves are characterised by difficulty and discrimination. The b parameter represents item difficulty as measured on a particular axis whose values range from -∞ to ∞; traditionally, the values are -3 to +3 when θ has a mean of 0 and a standard deviation of 1. Nevertheless, examinee ability over ±3 isn't common. Item difficulty value is high when the items are hard to answer. Low-ability examinees are less likely to get the correct answer. Easy items are those with low difficulty. Candidates with lower ability values are the potential to answer test items correctly. As for discrimination, that is also called "a" parameter. This information relates to whether an item may discriminate between examinees with abilities below and above the item location. The discriminating index parameter is calculated by tangentially connecting the curve to the difficulty level (b) parameter (Baker, 2001; Baker & Kim, 2017). A discriminating index parameter ranges from -∞ to ∞, with a typical value of ≤ 2.0. Hence, the steeper the curve is, the more discriminative the item is. (Baker, 2001; Bichi et al., 2019; Clark & Watson, 2019; Pliakos et al., 2019) indicates that low 'a' values are not useful for discriminating between ability levels. IRT's third parameter is guessing, called the 'c' parameter. Examinees of low ability respond correctly to an item when this parameter is lower than the asymptote parameter. When the three-parameter model is used, the parameter 'c' has the theoretical range of $0 \leq c \geq 1.0$, however, values above 0.35 are considered unacceptable (Ayanwale, 2019). Therefore, $\theta \leq c \leq 0.35$ is usually used.

4.2 Item response theory and its model

IRT items can be scored dichotomous or polytomous. Four models are usually used for dichotomous items. They are classified into one, two, three, and four-parameter logistic models (Cai et al., 2016; Cappelleri et al., 2014). However, the three most frequently used parameters are difficulty (b), discrimination (a), and guessing (c). Before each model can be used, it must meet all the necessary assumptions. The simplest of the three models, the one-parameter logistic model, is also called the Rasch model (Crocker & Algina, 1986; Hambleton & Swaminathan, 1985; Nataranjan, 2009). A logistic function between an examinee's ability (θ) and the difficulty of the question (b) is assumed to determine the chance that a correct answer will be provided. This is illustrated below.

$$P(\theta/b) = \frac{1}{1+\exp^{-1(\theta-b)}} \quad \text{.....Eqn. 2}$$

Furthermore, the two-parameter logistic model is a dichotomous IRT, in which the shape of the item response function is governed by two parameters, discrimination (a) and difficulty (b). The item response function increases monotonically when 'a' is positive or negative. However, as 'a' increases, the slope steepens. Positive item response functions are located with the larger value of 'b.' Examinees with the ability (θ) have the following likelihood of answering test items correctly. This is illustrated below.

$$P(\theta/a,b) = \frac{1}{1+\exp^{-a(\theta-b)}} \quad \text{.....Eqn. 3}$$

A parameter "c" calculates a lower asymptote parameter of the three-parameter logistic model, especially useful for multiple-choice and true-false tests. As c increases, the lower limit of the item response function also increases. The expression is as follows:

$$P(\theta/a,b,c) = C + (1-C)\frac{1}{1+\exp^{-a(\theta-b)}} \qquad (0 \le C \le 1) \quad \text{.....Eqn. 4}$$

Fourth parameter 'd' logistic models are dichotomous IRT models in which an upper asymptote parameter is added to the three-parameter model. As 'd' increases, the upper limit of the item response function (IRF) increases. Even with extreme levels of a trait, some items are so difficult that students cannot answer them all. The item's upper asymptote doesn't equal 1. The model fit will be improved by including a lower and an upper bound for the item response (Reise & Waller, 2009). A common use is to assess disorders that lead to extremely rare behavior. Hence, it is possible to expect that adding parameters will lead to an increasingly complex and well-fitting model. A mathematical expression for the model is:

$$P(\theta/a,b,c,d) = C + \frac{d-c}{1+\exp^{-a(\theta-b)}} \qquad (0 \le c < d \le 1) \quad \text{.....Eqn. 5}$$

Its advantage over CTT is that only adequately scored IRT can detect the significant differences between individuals whose scores are slightly different. When trait scores are incredibly high or low, they are out of the normal range. The IRT method solves this problem. Reise and Waller (2009) stipulate that items should be "difficult" enough for the levels of the trait in question. As a result, the four logistic parameter model, which incorporates time and slowness time responses, has yet to be fully integrated into conventional IRT models (Zhang, 2012).

4.3 Item response theory item analysis

The process of item analysis consists of assessing an item's quality in a test and the test as a whole based on the test results (Sim & Rasiah, 2006). In this way, items can be improved for future use, while those that are inadequate can be discarded. IRT analyses a scale at the item level by calculating item difficulty, discrimination, and test information function. Further, it computes the standard error (SE) for parameters "a" and "b" for each item and estimates the relationship between items and the constructs. Items may be positioned around theta (θ) or distributed uniformly from - ∞ to + ∞, depending on the purpose of the analysis. The location parameters of the instrument should be as close to the cut-off as possible when used to identify examinees for remedial measures or grouping them. For IRT models to be fully effective, item parameters must be calibrated with the right model.

The IRT model that best fits the data determines the model for item calibration of a test under development. An analysis of model-data fit is the only way to determine the right choice of item response theory models, as proposed by (Lee & Ansley, 2007). The model-data fit of item response theory models is critical when applied to real data. Estimated parameters may be compromised when a model does not fit the data (Bovaird & Embretson, 2012; Cai et al., 2016). To validate item response theory applications, fit tests of models need to be performed (González & Wiberg, 2017). According to Embretson and Reise (2013), checking item fit involves some issues. Item fit analyses can be used to identify a test model that retains the integrity of observed data, to identify extraneous dimensions that affect test item responses, and as a method of identifying faulty item construction, that is, incorrect keying and item fit, that is, those that indicate calibration errors during test development.

An item that does not fit a specific model is considered a poor fit (Hambleton & Jones, 1993). Comparing the observed performance of individual items with the predicted performance under the chosen model is a common way to assess model-data fit (Lee & Ansley, 2007; Yu et al., 2007). Based on Courville (2005), plots of observed and predicted score distributions or the chi-square test may be used to compare observed and predicted data. In Embretson and Reise (2013), examinees are first ranked according to their estimates (θ), then grouped into fixed or subjective categories. According to an item response function or item

characteristics curve, the proportion of examinees that answer an item correctly is calculated. A literature review on chi-square research shows that no chi-square fit index is preferred over another (Hambleton & Swaminathan, 1985). In Reise (1990), expressed chi-square as follows:

$$\chi^2 = \sum_{j=1}^{H} \frac{N_j \left(O_{ij} - E_{ij}\right)^2}{E_{ij}\left(1 - E_{ij}\right)} \quad \text{.....Eqn. 6}$$

'i' is the item, 'j' describes the interval based on examinees' ability estimates, 'H' represents the number of examinees within any interval, 'Nj' indicates the number of examinees with (θ) estimates within a given interval, and 'Eij, 'the expected proportion of keyed responses for intervals using an item response function evaluated at the median (θ) estimate within an interval. Chi-squares with high estimates diagnose items that do not fit the model, that is, those items performing differently than expected.

The likelihood ratio (G^2) is a chi-square statistic representing two tests of overall fit when items on a test are ten or less and twenty or more. (Rupp, 2003; Tuerlinckx et al., 2004) calculate the chi-square (χ^2) statistic as follows:

$$G^2 = 2\sum_{i=1}^{2^n} r_i \log_e \frac{r_i}{N - P_i} \quad i \quad \text{.....Eqn.7}$$

Where 2^n represents the number of possible patterns for each 'n' binary item scoring, 'r_i' is the observed frequency of pattern 'i', 'N' is the number of respondents, and 'P_i' is the estimated marginal probability. The number of degrees of freedom is 2^n-K_n-1, where K is the number of parameters in the response model. Thus, if 'G^2'>a critical value, the null hypothesis is rejected, and the ICC is expected to fit the item (Rupp, 2003).

4.4 Item and test information function

Test development and evaluation benefit from item information functions when lCC are fitted to test data. The corresponding item statistics and item information functions (IIF) will be incorrect if the ICCs do not fit the data well. It may be hard to use an item in all tests even when the fit is good if the parameter is low and the parameter is high. Additionally, an item may provide considerable information at one end of the ability continuum but be of no use on another end of the continuum. The information functions indicate how each item and the test estimates ability over the scale. IRT considers the test information function as a reliability coefficient since the variance measures the precision of measurement (Alagoz, 2005). Asymptotic distribution of the maximum likelihood estimator θˆ has mean θ and variance $\sigma^2 = \frac{1}{I(\theta)}$, where $I(\theta)$ is the amount of information. The ability estimate will be less precise, and the available information about an examinee's ability will be less when the variance of an estimator is large. The information function for the test with *n* items is defined as:

$I(\theta)=\sum_{i=1}^{n}\frac{[P_i(\theta)]^2}{P_i(\theta)Q_i(\theta)}$, Where $P_i(\theta)$ is obtained by evaluating the item characteristic curve at θ and $P_i(\theta)=\frac{\partial P_i}{\partial \theta}$. The item information is the decomposition of test information into each item. It is given as:

$I_i(\theta)=\sum_{i=1}^{n}\frac{[P'_i(\theta)]^2}{P_i(\theta)Q_i(\theta)}$, where $P_i'(\theta)=\frac{\partial P_i}{\partial \theta}$ ……..Eqn.8

Therefore, CTT's reliability coefficient and item reliability correspond to the test and IIF (Alagoz, 2005). An important element of IRT is the test information function (TIF). It shows how well the test estimates ability across a broad range of scores. A test is used to assess a person's ability so that the amount of information obtained from the test at any level can also be obtained (Birnbaum, 1968). In a test, there are items; therefore, the test information consists of the item information at a given level of ability. Each item contains a specific amount of information. The mathematical definition of item information may differ depending on the item characteristic curve model employed. The test information function therefore is:

$I(\theta) = \sum_{i=1}^{N} I_i(\theta)$ …….Eqn.9

$I(\theta)$ is the amount of information in a test at any ability level θ, $I_i(\theta)$ is the amount of information in each item, and N is the number of items in the test. Specifically, the TIF predicts the degree of accuracy at which we can measure any value of latent ability. Generally, the level of information in a test will be higher than that in a single-item test (Baker, 2001). When several items are included in a test, the greater the amount of information is revealed. More extended tests better measure test takers' abilities than shorter tests. A test information function may be used to balance multiple alternate test forms for the same exam. TIF values should be the same across all alternate forms (Song et al., 2019).

5. Conclusion and Recommendations

The present article discusses CTT and IRT in ECOL's test development and item analysis. Educational assessment includes the performance of tests; their results are used to inform various educational decisions. Tests are therefore widely regarded as an important part of education. Testing is a method of evaluating a candidate's ability in a previously defined knowledge or skill domain. To better understand the relationship between the observed (or actual) score on an examination and the unobserved proficiency in the domain, we need a test theory model. CTT and IRT are commonly used models. The CTT calculates statistics such as correlations among items, covariance's, difficulties, discrimination power, reliability coefficients, variance/standard deviation of the sample, measurement errors, etc., to improve the reliability and validity of measurement tools. The theory deals with important measurement problems from a constant perspective. Due to several weaknesses of CTT, the need for another test theory emerged. These include item and test statistics that differed

across tests and groups; a single error estimate was produced for individuals of all skills levels, and the weakness in test equating. A significant innovation in educational assessment and psychometrics has been the development of IRT. Models of IRT have been used extensively in test development and assessment over the past several decades, attesting to their importance. The IRT models analyse items, assemble test forms, and equate. Despite being helpful in many situations, IRT models use strong assumptions and are mathematically more complex than CTT models used in ECOL. In conclusion, the study strongly recommends that ECOL shift its test development and item analysis modus operandi from CTT to modern test theory, which has numerous benefits.

Conflict of interest
The authors declare no conflict of interest.

Acknowledgment
The authors would like to thank the reviewers for their time and efforts in reviewing the manuscript. The manuscript has been improved by all valuable comments and suggestions, we are deeply grateful.

6. References

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

Adedoyin, O. O. (2010). Investigating the Invariance of Person Parameter Estimates Based on Classical Test and Item Response Theories. *International Journal of Educational Sciences*, *2*(2), 107–113. https://doi.org/10.31901/24566322.2010/02.02.07

Adegoke, B. A. (2013). Comparison of item statistics of physics achievement test using Classical test theory and item response theory frameworks. *Journal of Education and Practice*, *22*(4), 87–96. www.iiste.org

Adewale, J.G., Adegoke, B.A., Adeleke, J.O. & Metibemu, M. A. (2017). *A Training Manual On Item Response Theory*. Institute of Education, University of Ibadan in Collaboration with National Examinations Council, Minna, Niger State.

Alagoz, C. (2005). *Scoring tests with dichotomous and polytomous items*. https://getd.libs.uga.edu/pdfs/alagoz_cigdem_200505_ma.pdf

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

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

Ayanwale, Musa Adekunle, Adeleke, J. O., & Mamadelo, T. I. (2019). Invariance Person Estimate of Basic Education Certificate Examination: Classical Test Theory and Item Response Theory Scoring Perspective. *Journal of the International Society for Teacher Education*, *23*(1), 18–26. https://files.eric.ed.gov/fulltext/EJ1237578.pdf

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

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

54205-8_1

Behavior, S., Yen, Y. C., Chen, H., & Cheng, M. (2012). The Four-Parameter Logistic Item Response Theory Model As a Robust Method of Estimating Ability Despite Aberrant Responses. *Social Behavior and Personality: An international journal*, 40(10), 1679-1694.https://doi.org/10.2224/sbp.2012.40.10.1679

Bichi, A. A., Embong, R., Talib, R., Salleh, S., & Bin Ibrahim, A. (2019). Comparative Analysis of Classical Test Theory and Item Response Theory using Chemistry Test Data. *International Journal of Engineering and Advanced Technology*, *8*(5), 1260–1266. https://doi.org/10.35940/ijeat.E1179.0585C19

Bichi, A. A., & Talib, R. (2018). Item Response Theory: An Introduction to Latent Trait Models to Test and Item Development. *International Journal of Evaluation and Research in Education*, *7*(2), 142. https://doi.org/10.11591/ijere.v7i2.12900

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In: Lord, F.M. and Novick, M.R., Eds., Statistical Theories of Mental Test Scores, Addison-Wesley, Reading, 397-479.

Bovaird, J. A., & Embretson, S. E. (2012). Modern Measurement in the Social Sciences. In *The SAGE Handbook of Social Research Methods* (pp. 268–289). SAGE Publications Ltd. https://doi.org/10.4135/9781446212165.n16

Brown, J. D. (2013). Classical test theory. In *The Routledge Handbook of Language Testing* (pp. 323–335). Springer, Singapore. https://doi.org/10.4324/9780203181287-35

Cai, L., Choi, K., Hansen, M., & Harrell, L. (2016). Item Response Theory. In *Annual Review of Statistics and Its Application* (Vol. 3, pp. 297–321). Annual Reviews Inc. https://doi.org/10.1146/annurev-statistics-041715-033702

Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics*, *36*(5), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006

Chen, W. H., & Thissen, D. (1997). Local Dependence Indexes for Item Pairs Using Item Response Theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. https://doi.org/10.3102/10769986022003265

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412–1427. https://doi.org/10.1037/pas0000626

Cohen, R.J., & Swerdlik, M. E. (2009). *Psychological testing and assessment: An introduction to tests and measurement.* (4th ed.). Mayfield Publishing House.

Cohen, R. . J., Swerdlik, M. E., & Sturman, E. (2013). Psychological testing and assessment : an introduction to tests and measurement. *Psychological Assessment*, *53*(4), 55–67. https://perpus.univpancasila.ac.id/repository/EBUPT181396.pdf

Courville, T. G. (2005). An empirical comparison of item response theory and classical test theory item/person statistics. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, *65*(7), 2575. https://oaktrust.library.tamu.edu/handle/1969.1/1064

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Harcourt Brace Jovanovich. https://eric.ed.gov/?id=ED312281

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. In *Medical Education,* 44(1), 109–117. https://doi.org/10.1111/j.1365-2923.2009.03425.x

Debelak, R., & Koller, I. (2020). Testing the Local Independence Assumption of the Rasch Model With Q3-Based Nonparametric Model Tests. *Applied Psychological Measurement*, *44*(2), 103–117. https://doi.org/10.1177/0146621619835501

Demars, C. E. (2017). Classical test theory and item response theory. In *The Wiley*

Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development, 2(1), 49–73. https://doi.org/10.1002/9781118489772.ch2

Dent, J. A., Harden, R. M., & Hunt, D. (2001). A Practical Guide for Medical Teachers. *Journal of the Royal Society of Medicine*, *94*(12), 653–653. https://doi.org/10.1177/014107680109401222

DeVellis, R. F. (2006). Classical test theory. *Medical Care*, *44*(11), 50-59. https://doi.org/10.1097/01.mlr.0000245426.10853.30

Downing, S. M. (2003). Item response theory: Applications of modern test theory in medical education. *Medical Education*, *37*(8), 739–745. https://doi.org/10.1046/j.1365-2923.2003.01587.x

Ebel, R. L. (1965). Book Reviews : Measuring Educational Achievement. *Educational and Psychological Measurement*, *25*(4), 1167–1169. https://doi.org/10.1177/001316446502500428

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*(1), 5–18. https://doi.org/10.1007/s11136-007-9198-0

Elgadal, A. H., & Mariod, A. A. (2021). Item Analysis of Multiple-choice Questions (MCQs): Assessment Tool For Quality Assurance Measures. *Sudan Journal of Medical Sciences*, 16(3), 334-346. https://doi.org/10.18502/sjms.v16i3.9695

Embretson, S. E., & Reise, S. P. (2013). Item Response Theory for Psychologists. Lawrence Erlbaum Associates, Inc., Mahwah. 1–371. https://doi.org/10.4324/9781410605269

Esmaeeli, B., Shandiz, E. E., Norooziasl, S., & Shojaei, H. (2021). The Optimal Number of Choices in Multiple-Choice Tests : A Systematic Review. *Med Edu Bull*, *2*(5), 253–260. https://doi.org/10.22034/MEB.2021.311998.1031

Exam Council of, L. (2018). *Establishment of ECOL*. https://www.google.com/search?q=Exam+Council+of+Lesotho%2C+2018&oq=Exam+Council+of+Lesotho%2C+2018&aqs=chrome..69i57j33i160l2.2034j0j7&sourceid=chrome&ie=UTF-8

Filgueiras, A., Hora, G., Fioravanti-Bastos, A. C. M., Santana, C. M. T., Pires, P., De Oliveira Galvão, B., & Landeira-Fernandez, J. (2014). Development and psychometric properties of a novel depression measure. *Temas Em Psicologia*, *22*(1), 249–269. https://doi.org/10.9788/TP2014.1-19

Finch, H., & Monahan, P. (2008). A bootstrap generalization of modified parallel analysis for IRT dimensionality assessment. *Applied Measurement in Education*, *21*(2), 119–140. https://doi.org/10.1080/08957340801926102

Finch, W. H., & French, B. F. (2015). Modeling of Nonrecursive Structural Equation Models With Categorical Indicators. *Structural Equation Modeling*, *22*(3), 416–428. https://doi.org/10.1080/10705511.2014.937380

Ganglmair, A., & Lawson, R. (2010). Advantages of Rasch modelling for the development of a scale to measure affective response to consumption. In *E-European Advances in Consumer Research, 6,* 162–167. https://www.acrwebsite.org/volumes/11738

Gay, L.R, Miles, G. E. & Airasian, P. (2011). Educational Research: Competencies for Analysis and Applications. 10th Edition, Pearson Education International, Boston.

González, J., & Wiberg, M. (2017). Applying Test Equating Methods using R. *Methodology of Educational Measurement and Assessment*. https://link.springer.com/bfm:978-3-319-51824-4/1

Hambleton, R.K. and Swaminathan, H. (1985). *Item response theory: principles and applications.* p.332. https://doi.org/10.1177/014662168500900315

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hill, C., Nel, J. A., van de Vijver, F. J. R., Meiring, D., Valchev, V. H., Adams, B. G., & de Bruin, G. P. (2013). Developing and testing items for the South African Personality Inventory. *SA Journal of Industrial Psychology*, *39*(1), 1-13. https://doi.org/10.4102/sajip.v39i1.1122

Hingorjo, M. R., & Jaleel, F. (2012). Analysis of one-best MCQs: The difficulty index, discrimination index and distractor efficiency. *Journal of the Pakistan Medical Association*, *62*(2), 142–147. https://pubmed.ncbi.nlm.nih.gov/22755376/

Immekus, J. C., Snyder, K. E., & Ralston, P. A. (2019). Multidimensional Item Response Theory for Factor Structure Assessment in Educational Psychology Research. *Frontiers in Education*, *4*. https://doi.org/10.3389/feduc.2019.00045

IResearchNet (2022). *Classical Test Theory*. http://psychology.iresearchnet.com/industrial-organizational-psychology/i-o-psychology-theories/classical-test-theory/

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment. *Applied Psychological Measurement*, *40*(8), 559–572. https://doi.org/10.1177/0146621616664046

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, *30*(1), 17–24. https://doi.org/10.1037/h0057123

Khan, H. F., Danish, K. F., Awan, A. S., & Anwar, M. (2013). Identification of technical item flaws leads to improvement of the quality of single best multiple choice questions. *Pakistan Journal of Medical Sciences*, *29*(3), 715. https://doi.org/10.12669/pjms.293.2993

Kim, D., de Ayala, R. J., Ferdous, A. A., & Nering, M. L. (2011). The comparative performance of conditional independence indices. *Applied Psychological Measurement*, *35*(6), 447–471. https://doi.org/10.1177/0146621611407909

Kline, R. B. (2005). "*Principles and practice of structural equation modelling "*. ((2nd ed.)). The Guilford Press.

Kline, T. (2014). Classical Test Theory: Assumptions, Equations, Limitations, and Item Analyses. In *Psychological Testing: A Practical Approach to Design and Evaluation*, 23(2), 91–106. https://doi.org/10.4135/9781483385693.n5

Kolen, M. J. (1981). Comparison of traditional and Item Response Theory methods for equatingTests. *Journal of Educational Measurement*, *18*(1), 1–11. https://doi.org/10.1111/j.1745-3984.1981.tb00838.x

Krishnan, V. (2013). The Early Child Development Instrument (EDI): An item analysis using Classical Test Theory (CTT) on Alberta ' s data. Early Child Development Mapping (ECMap) Project Community-University Partnership (CUP) Faculty of Extension, University of Alberta.

Lang, J. W. B., & Tay, L. (2021). The Science and Practice of Item Response Theory in Organizations. In *Annual Review of Organizational Psychology and Organizational Behavior*, 8, 311–338. https://doi.org/10.1146/annurev-orgpsych-012420-061705

Lee, W., & Ansley, T. N. (2007). Assessing IRT Model-Data Fit for mixed format tests. *Journal of Applied Psychology*, 92(2), 23–50. http://dx.doi.org/10.1026/apl0000636

Lord, F. M. (2012). Applications of item response theory to practical testing problems. In *Applications of Item Response Theory To Practical Testing Problems*. https://doi.org/10.4324/9780203056615

Magis, D. (2007). *Influence, Information and Item Response Theory in Discrete Data Analysis.* Retrieved on 12 June, 2022 from http://bictel.ulg.ac.be/ETDdb/collection/available/ULgetd-06122007-100147/.

Mona, N. (2014). Application of Classical Test Theory and Item Response Theory to Analyze Multiple Choice Questions (Unpublished doctoral thesis). University of Calgary, Calgary, AB. doi:10.11575/PRISM/24958

Nataranjan, V. (2009). *Basic Principle of Item Response Theory and Application to Practical Testing and Assessement.* Merit Trac Services Publishing Ltd.

Ojerinde, D. & Ifewulu, B. C. (2012). Item Unidimensionality Using 2010 Unified Tertiary Matriculation Examination Mathematics Pre-test. *A Paper Presented at the 2012 International Conference of IAEA*, 5–18.

Pliakos, K., Joo, S. H., Park, J. Y., Cornillie, F., Vens, C., & Van den Noortgate, W. (2019). Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems. *Computers and Education*, *137*, 91–103. https://doi.org/10.1016/j.compedu.2019.04.009

Preston, R., Gratani, M., Owens, K., Roche, P., Zimanyi, M., & Malau-Aduli, B. (2020). Exploring the Impact of Assessment on Medical Students' Learning. *Assessment and Evaluation in Higher Education*, *45*(1), 109–124. https://doi.org/10.1080/02602938.2019.1614145

Privitera, G. J. (2012). *Statistics for the behavioral sciences.* Sage Publications, Inc. https://psycnet.apa.org/record/2011-21294-000

Reckase, M. D. (2009). *Multidimensional Item ResponseTheory*. Springer Verlag.

Reise, S. P. (1990). A Comparison of Item- and Person-Fit Methods of Assessing Model-Data Fit in IRT. *Applied Psychological Measurement*, 14(2), 127–137. https://doi.org/10.1177/014662169001400202

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27–48. https://doi.org/10.1146/Annurev.Clinpsy.032408.153553

Rupp, A. A. (2003). Item Response Modeling With BILOG-MG and MULTILOG for Windows. *International Journal of Testing*, *3*(4), 365–384. https://doi.org/10.1207/s15327574ijt0304_5

Rusch, T., Lowry, P. B., Mair, P., & Treiblmaier, H. (2017). Breaking free from the limitations of classical test theory: Developing and measuring information systems scales using item response theory. *Information and Management*, 54(2), 189–203. https://doi.org/10.1016/j.im.2016.06.005

Sim, S. M., & Rasiah, R. I. (2006). Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. *Annals of the Academy of Medicine Singapore*, *35*(2), 67–71. http://www.ams.edu.sg

Song, Y., Kim, H., & Park, S. Y. (2019). An Item Response Theory Analysis of the Korean Version of the CRAFFT Scale for Alcohol Use Among Adolescents in Korea. *Asian Nursing Research*, *13*(4), 249–256. https://doi.org/10.1016/j.anr.2019.09.003

Steyer, R. (2001). Classical (Psychometric) Test Theory. *International Encyclopedia of the Social & Behavioral Sciences*, 1955–1962. https://doi.org/10.1016/B0-08-043076-7/00721-X

Tang, X., Karabatsos, G., & Chen, H. (2020). Detecting Local Dependence: A Threshold-Autoregressive Item Response Theory (TAR-IRT) Approach for Polytomous Items. *Applied Measurement in Education*, 280–292. https://doi.org/10.1080/08957347.2020.1789136

Tay, L., Meade, A. W., & Cao, M. (2015). An Overview and Practical Guide to IRT Measurement Equivalence Analysis. *Organizational Research Methods*, *18*(1), 3–46.

https://doi.org/10.1177/1094428114553062

Toksöz, S., & Ertunç, A. (2017). Item Analysis of a Multiple-Choice Exam. *Advances in Language and Literary Studies*, *8*(6), 141. https://doi.org/10.7575/aiac.alls.v.8n.6p.141

Traub, R. E. (2015). Classical test theory in historical perspective. *Journal of Educational Measurement: Issues and Practice*, *16*(4), 8–14. https://doi.org/10.1111/emip.2015.16.issue-4

Tuerlinckx, F., Rijmen, F., Molenberghs, G., Verbeke, G., Briggs, D., Van den Noortgate, W., Meulders, M., & De Boeck, P. (2004). Estimation and software. In *Explanatory Item Response Models*, 6, 343–373. https://doi.org/10.1007/978-1-4757-3990-9_12

Vyas, R., & Supe, A. (2008). Multiple choice questions: A literature review on the optimal number of options. In *National Medical Journal of India*, 21(3), 130–133. https://pubmed.ncbi.nlm.nih.gov/19004145/

Wells, C. S., & Wollack, J. A. (2018). An Instructor's Guide to Understanding Test Reliability. *Testing and Evaluation Services*, 1–7. https://testing.wisc.edu/Reliability.pdf

Yen, W. M. (1993). Scaling Performance Assessments: Strategies for Managing Local Item Dependence. *Journal of Educational Measurement*, *30*(3), 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Yu, C. H., Popp, S. O., Digangi, S., & Jannasch-Pennell, A. (2007). Assessing unidimensionality: A comparison of Rasch modeling, Parallel analysis, and TETRAD. *Practical Assessment, Research and Evaluation*, *12*(14), 1–19. https://doi.org/https://doi.org/10.7275/q7g0-vt50

Zhang, J. (2012). Calibration of Response Data Using MIRT Models With Simple and Mixed Structures. *Applied Psychological Measurement*, *36*(5), 375–398. https://doi.org/10.1177/0146621612445904

Zhu, X., & Lu, C. (2017). Re-evaluation of the New Ecological Paradigm scale using item response theory. *Journal of Environmental Psychology*, 54, 79–90. https://doi.org/10.1016/j.jenvp.2017.10.005