Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 29 Number 14, November 2024 ISSN 1531-7714

From Investigating the Alignment of *A Priori* Item Characteristics Based on the CTT and Four-Parameter Logistic (4-PL) IRT Models to Further Exploring the Comparability of the Two Models

Agus Santoso, *Universitas Terbuka*
Heri Retnawati, *Universitas Negeri Yogyakarta*
Timbul Pardede, *Universitas Terbuka*
Ezi Apino, *Universitas Negeri Yogyakarta*
Ibnu Rafi, *Universitas Negeri Yogyakarta*
Munaya Nikma Rosyada, *Universitas Negeri Yogyakarta*
Gulzhaina K. Kassymova, *Abai Kazakh National Pedagogical University*
Xu Wenxin, *Abai Kazakh National Pedagogical University*

The test blueprint is important in test development, where it guides the test item writer in creating test items according to the desired objectives and specifications or characteristics (so-called *a priori* item characteristics), such as the level of item difficulty in the category and the distribution of items based on their difficulty level. Given that the difficulty level of the test items (easy, medium, or hard) created is influenced by the perceptions, knowledge, and experience of the item writer, item analysis based on empirical data using a specific measurement framework needs to be conducted, in addition to evaluation based on expert judgment, to ensure that the test items and the test itself have appropriate characteristics. The present study investigated the extent to which the *a priori* characteristics (i.e., item difficulty) of the items of the Business English test taken by 4,836 Universitas Terbuka (UT) students aligned with their characteristics when estimated under classical test theory (CTT) and four-parameter logistic (4-PL) IRT models based on empirical data. In light of the two measurement models used, CTT and 4-PL, we extended this study to exploring the comparability of the two models based on the yielded item difficulty and discrimination estimates and the relationship between pseudo-guessing and carelessness parameters. Our study suggested insufficient support for asserting that the characteristics of the items used in the Business English test align with the characteristics expected by the test developers. The exploration of the comparability of the CTT and 4-PL models demonstrated that while the two models were comparable in terms of the item difficulty estimates yielded, they were not comparable for the item discrimination estimates. Our study also did not find a linear association of the pseudo-guessing and carelessness parameters estimated under the 4-PL model. Further findings of our study and their implications, especially on test development practices, are discussed.

Keywords: carelessness, classical test theory, item characteristics, item response theory

Introduction

Educational measurement and assessment require quality instruments to obtain accurate results. The test is one of the most frequently used instruments in educational measurement. It is widely recognized that objectivity, validity, and reliability are fundamental criteria that determine the quality of a test (AERA et al., 2014; Cohen & Swerdlik, 2018; Crocker & Algina, 2008; Ebel & Frisbie, 1991; Miller et al., 2009; Mohajan, 2017), in addition to other criteria that also deserve attention such as the usefulness, fairness (McMillan, 2000), and cost efficiency of administering, scoring, analyzing the results of scoring, and interpreting the results of the analysis (Cohen & Swerdlik, 2018). A series of stages should be followed by test developers to produce a quality test, starting from conceptualizing the test, constructing test items, piloting the test, analyzing test items, to developing guidelines on test administration, scoring, and interpretation (Cohen & Swerdlik, 2018; Crocker & Algina, 2008). Item analysis in test development is also deemed to be essential in assisting test developers to ensure that test items have the desired characteristics or psychometric properties.

The characteristics of test items that are frequently of concern and identified through quantitative item analysis in test development are item difficulty, item discriminating power, and the functioning or plausibility of distractors when test items are multiple-choice type (Lahza et al., 2023; Odukoya et al., 2018; Quaigrain & Arhin, 2017; Rafi et al., 2023; Ulwatunnisa et al., 2023; Yim et al., 2024). Quantitative item analysis was conducted using the empirical response data of test takers involved in the pilot stages. This analysis allows test developers to ensure the quality (Santoso, Pardede, Djidu, et al., 2022) and feasibility of test items and the overall test that will be used in the desired setting or context by revising or even eliminating certain test items that do not meet the required specifications. Moreover, quantitative item analysis allows test developers to provide empirical evidence on the difficulty level of the test items that test developers create based on what they perceive. This analysis, in other words, offers usefulness in the sense that it ensures that the number of easy, medium, and hard items on the test is in the expected proportion, which is usually adjusted according to the purpose of the test. In constructing the test items, test developers need to adhere to the specifications defined for the test, where the specifications are expressed in a test blueprint (or test specifications) where one of the aspects regulated in the test blueprint is the level of item difficulty (Sayin & Bulut, 2024). The aspect of the difficulty level of the test items contained in the test blueprint plays a role in ensuring that the items to be administered to test takers can cover all levels of ability of test takers.

Classical test theory (CTT) and item response theory (IRT) are two main frameworks that have their own advantages and disadvantages that can be used in quantitative item analysis to determine characteristics of test items in educational measurement. The main advantage of CTT is its simplicity (Hu et al., 2021; Progar & Sočan, 2008), but the estimation of item statistics which relies heavily on sample characteristics is considered to be the main drawback of CTT (Ayanwale et al., 2018; Baker & Kim, 2017; Kartowagiran et al., 2019; Hambleton et al., 1991; Santoso, Pardede, Apino, et al., 2022; Setiawati et al., 2023). On the other hand, IRT is a model-based paradigm; it starts by modeling the relationship between the latent variable being measured and the item responses (Baker & Kim, 2017; DeMars, 2010; Hambleton et al., 1991; Progar & Sočan, 2008). An important feature offered by the IRT modeling approach, which is also recognized as its main advantage, is that the person parameter is independent of the item parameter. IRT, however, requires more rigorous assumptions than CTT (Eleje et al., 2018; Hu et al., 2021), more complex statistical calculations (Eleje et al., 2018; Jian et al., 2021; Kalkan & Çuhadar, 2020), and a larger sample size (Eleje et al., 2018; Hu et al., 2021).

In educational measurement practice, those two theories or frameworks, CTT and IRT, have been used simultaneously to describe the characteristics of test items (Adegoke, 2013; Ayanwale et al., 2018; Subali et al., 2021). However, one of the issues related to the use of CTT and IRT together in quantitative item analysis is whether the two are comparable (Awopeju & Afolabi, 2016; Eleje et al., 2018; Progar & Sočan, 2008; Setiawati et al., 2023). Previous studies (Awopeju & Afolabi, 2016; Bichi et al., 2019; Eleje et al., 2018; Magno, 2009; Progar & Sočan, 2008; Setiawati et al., 2023) have been conducted to investigate test item characteristics (especially in terms of item difficulty and discrimination) by using both CTT and IRT. However, inconsistent results have been identified in those previous studies.

Comparative analysis between CTT and IRT has attracted the interest of many researchers. Progar and Sočan (2008) conducted an empirical comparison between CTT and IRT using a data set from the Trends in International Mathematics and Science Study 1995 (TIMSS 1995). They focused on investigating whether the item characteristics estimated using CTT and 2-PL IRT are comparable. In addition, Magno (2009) investigated the differences in the results of item characteristics based on the CTT and IRT approaches, in which the IRT model used was Rasch. Awopeju and Afolabi (2016) compared the results of item characteristics estimation using CTT and IRT on the senior school certificate mathematics examination. This study compared the estimation of item difficulty based on CTT and 1-PL IRT and compared the estimation of the discrimination parameter based on CTT and 2-PL IRT. Eleje et al. (2018) conducted a comparative study to compare the estimation results of item characteristics from the Diagnostic Quantitative Economics Skills Test using CTT and IRT. Although they used the 3-PL IRT model to estimate item parameters, only the parameters of difficulty and discrimination received more attention from them. Furthermore, Ayanwale et al. (2018) compared the results of estimation of item difficulty and discrimination from the Basic Education Certificate Examination using CTT and 3-PL IRT models. Besides, using Chemistry test data, Bichi et al. (2019) compared the results of item parameter estimation between CTT and IRT. Based on the model fit analysis, this study used the 2-PL IRT model to estimate item parameters, then the results were compared with the estimation results based on CTT. Most recently, Setiawati et al. (2023) conducted a study of the item parameters of the differential aptitude test using CTT and IRT. They compared the results of item characteristic estimation based on CTT, Rasch, 1-PL IRT, 2-PL IRT, and 3-PL IRT models. Based on previous studies, it is clear that the item parameter estimates with CTT have been compared with various dichotomous IRT models (i.e., 1-PL, 2-PL, and 3-PL), including the Rasch model. However, currently no studies have been found that compare test item characteristics based on CTT and 4-PL IRT model. The present study, therefore, seeks to fill that gap.

Existing literature reveals that the 4-PL IRT model is still less popular than other IRT models (Kalkan & Çuhadar, 2020; Loken & Rulison, 2010), but now 4-PL IRT is increasingly in demand (Barnard-Brak et al., 2018; Kalkan & Çuhadar, 2020; Liao et al., 2012; Primi, 2018; Robitzsch, 2022). The limited literature regarding the application of the 4-PL IRT model in educational measurement coupled with studies focusing on comparing the results of item parameter characteristic estimation between CTT and 4-PL IRT models are still scarce have motivated us to conduct this study. These two models also offer the opportunity to investigate the extent of support that empirical test-taker response data provide for the difficulty level of test items on a test blueprint or test specification when the empirical data are analyzed under the CTT and 4-PL IRT models. This study, therefore, has a two-fold objective: to reveal the alignment of the difficulty levels of test items set on a test blueprint or test specification (so-called *a priori* item characteristics) with those estimated under the CTT and 4-PL IRT models based on test taker response data and to reveal the comparison between the two models based on the estimated test item characteristic results. This study is expected to enrich the literature on this topic and provide insight to educators, practitioners, and test developers regarding the item calibration process using an appropriate framework.

Theoretical Foundation and Literature Review

This section discusses the two main theories that underlie this study in estimating test item characteristics, namely classical test theory (CTT) and item response theory (IRT). The discussion regarding IRT focuses on the 4-PL IRT model because this model estimates student abilities and test item characteristics by considering the anomalous behavior of low-ability and high-ability students in responding to test items.

Classical Test Theory (CTT)

Classical test theory (CTT) or true score theory is a measurement framework that makes it possible for researchers to understand, manipulate, and interpret measurement results. The CTT assumption is that every measurement is subject to error and that every observation is imperfect. CTT model decomposes the observed score from the measuring instrument into the true score and the error component. CTT model is mathematically expressed as follows (Equation 1).

$$X = T + E \quad (1)$$

where X is the observed measurement score or test score, T is the true (latent) measurement score or total test score; and E is random error. To use the CTT model, there are four additional assumptions beyond the general form presented in Equation 1: (1) $E(X) = T$, the expected value of the observed score is the actual score, (2) $\text{Cov}(T, E) = 0$, true scores and errors are independent, (3) $\text{Cov}(E1, E2) = 0$, errors in all test forms are independent, and (4) $\text{Cov}(E1, T2) = 0$, errors in one form of test are independent of actual scores on other forms of testing (Desjardins & Bulut, 2018). Because of this assumption, the CTT model can be represented as the sum of the orthogonal (i.e., uncorrelated) variance components as follows (Equation 2).

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 \quad (2)$$

Equation 2 states that the observed score variance is the sum of the actual score variance and the error variance. In this model, the variance of true scores is assumed to be constant (never changing regardless of instrument form, date of assessment, etc.), while the variance of errors fluctuates (e.g., some forms of a test may contain more errors than other forms of a test). Measurement errors can be divided into two types: random (i.e., unpredictable and inconsistent) and systematic (i.e., constant and predictable) errors (Desjardins & Bulut, 2018).

In the CTT model, item statistics are used to describe the characteristics of a test item, namely item difficulty and item discrimination. Both of these statistics can be applied to dichotomous and polytomous data, but in this study we focus on their application to dichotomous data. Item difficulty index is useful for evaluating whether the level of difficulty of an item corresponds to the ability level of the test takers (Allen & Yen, 1979). The item discrimination index indicates the degree to which responses to one item are related to responses to other items in the test (Allen & Yen, 1979). In other words, the item discrimination index indicates whether an item differentiates between test takers who perform well and those who perform poorly on the test. By considering the level of difficulty and item discrimination, the test developer is expected to be able to develop a test that can provide as much information as possible about the differences in the test takers on the trait being measured.

The item difficulty for item i, denoted by p_i, is defined as the proportion of test takers who answered item i correctly. Although the proportion of test takers who answered an item correctly has traditionally been named item difficulty, this proportion is logically more accurately described as item ease, because the proportion increases as items become easier (Allen & Yen, 1979). If p_i is close to 0 or 1, the item should be modified or discarded (Allen & Yen, 1979), as it provides no information about differences between test takers' trait or ability levels. If $p_i = 0$, no participant has the correct answer; this item was too difficult and completely useless. If $p_i = 1$, all participants answered correctly, this also does not provide information regarding differences in the nature or abilities of the test takers. An item will offer the maximum amount of information regarding differences among test takers when $p_i = 0.5$ (Allen & Yen, 1979; Crocker & Algina, 2008). However, the use of this suggestion is influenced by the intercorrelation between items. If all items are perfectly correlated with each other and have a difficulty of 0.5, half of the test takers will obtain a total test score of 0, and the other half will receive a perfect total test score. It will also not demonstrate good discrimination between the test taker's trait levels. Therefore, it is best to select items with an average difficulty range of around 0.5 (Allen & Yen, 1979; Rafi et al., 2023; Reynolds et al., 2009). Generally, item difficulty of around 0.3 to 0.7 maximizes the information the test provides about differences among test takers (Allen & Yen, 1979; Rafi et al., 2023). Following these guidelines, in this study, we defined an item with a p of less than 0.3 as "hard" and more than 0.7 as "easy".

Apart from item difficulty, another statistic that is used to describe item characteristics is item discrimination. The item discrimination index for items i, denoted by d_i, is calculated by the following formula (Equation 3) (Allen & Yen, 1979).

$$d_i = \frac{U_i}{n_{iU}} - \frac{L_i}{n_{iL}} \quad (3)$$

where U_i and L_i are the number of test takers who have total test scores above and below the range of total test scores and who also have item i correct, respectively; n_{iU} and n_{iL} are the number of examinees who have total scores test above and below the total test score range, respectively. Equation 3 shows that d_i is the difference between the proportion of test takers who scored high when answering the test items and the proportion of test takers who scored low when answering the test

items. The upper and lower ranges are generally defined as the 10% to 33% of the upper and lower samples, with test takers ordered by their total test scores (Allen & Yen, 1979). If the total test scores are normally distributed, it is optimal to use the 27% of test takers with the highest total test score as the upper range and use 27% of the test takers with the lowest total test score as the lower range. When total test scores are normally distributed, using the upper and lower 27% yields the best estimate of d_i (Allen & Yen, 1979). An alternative way to determine the item discrimination index is to use the point-biserial correlation, r_{iX}, between the score on item i and the total test score, X (see Equation 4).

$$r_{iX} = \frac{\overline{X}_i - \overline{X}}{s_X} - \sqrt{\frac{p_i}{1-p_i}} \qquad (4)$$

where $\overline{X}_i$ is the average X score among test takers who correctly answered item i, $\overline{X}$ and s_X are the mean and standard deviation of X scores among all test takers, and p_i is the difficulty level of the item i.

The item discrimination indices, d_i and r_{iX}, are valuable pieces of information in item analysis. Ideally, d_i and r_{iX} should be positive, indicating that more test takers with high scores than examinees with low scores were able to answer the item correctly. Items with negative values for d_i and r_{iX} seem to measure the opposite of what the test measures. A negative d_i or r_{iX} value may indicate that there is an error in the answer key, or the item has a poor redaction (Allen & Yen, 1979). Items with low or negative d_i or r_{iX} should be revised or eliminated (Allen & Yen, 1979; Hopkins, 1998; Reynolds et al., 2009). Hopkins (1998) provides a guide for evaluating item discrimination. Items with $d \leq 0.10$ are considered "poor", $0.10 < d < 0.30$ are considered "fair", and $d \geq 0.30$ are considered "good" in discriminating the performance of test takers (Hopkins, 1998; Reynolds et al., 2009). In this study, we decided to follow these guidelines to evaluate item discrimination.

The Four-Parameter Logistic Item Response Theory Model (the 4-PL IRT Model)

IRT models for dichotomous response data usually assume a logistic curve for the probability of the 'correct answer' as a function of the underlying latent construct (θ). In IRT, there are three popular parameter logistic (PL) models: 1-PL, 2-PL, and 3-PL (Baker, 2001; Baker & Kim, 2017; Hambleton et al., 1991). The 1-PL model assumes that all items have the same slope (item discrimination) and only differ in terms of location (item difficulty). The 2-PL model assumes that apart from having different locations, items may also have different slopes. In 1-PL and 2-PL models, the probability of answering correctly ranged between 0 and 1. The probability of answering correctly was close to 0 in the case of low ability participants correctly answering difficult items, and 1 in the case of high ability participants correctly answering easy items. This assumption may not always be correct because clueless test takers may choose the correct answer by guessing (Liao et al., 2012). In addition, in multiple-choice tests, the probability of answering correctly may not be close to 0 even for participants with low ability (Barton & Lord, 1981; Liao et al., 2012). To overcome this condition, Birnbaum (2008) introduced a lower asymptote to model a situation where test takers make random guesses in answering an item. This model is known as the 3-PL model. The 3-PL model, besides assuming the items may have different locations and slopes, this model also predicts the lower asymptotes (or pseudo-guessing). In the 3-PL model, when the lower asymptote is set equal to 0, we obtain the 2-PL model.

In other conditions, high-ability test takers sometimes incorrectly answered items they should have answered correctly when they were anxious, careless, distracted by poor test conditions, or when they misread the questions (Kalkan & Çuhadar, 2020; Liao et al., 2012; Loken & Rulison, 2010; Robitzsch, 2022). In this condition, the 3-PL model may be very detrimental for high ability participants who make careless mistakes on easy items that they should be able to answer correctly (Barton & Lord, 1981; Kalkan & Çuhadar, 2020; Liao et al., 2012; Loken & Rulison, 2010). Moreover, in the 3-PL model, the lower asymptote accommodates a situation where the low-ability test taker correctly guesses the difficult item, but the upper asymptote with a value of 1 gives a probability of 0 for the high-ability participant failing to answer the easy item correctly (Liao et al., 2012; Loken & Rulison, 2010). To overcome this condition, the 4-PL model is considered fairer. Barton and Lord (1981) introduced the upper asymptote parameter, denoted by u, into the 3-PL model, resulting in a 4-PL model which is mathematically expressed as follows (Equation 5).

$$\mathrm{P}\left(X_{ij}=1 \mid \theta_i; a_j, b_j, c_j, u_j\right) = c_j + \left(u_j - c_j\right)\frac{\exp\left[1.7a_j\left(\theta_i - b_j\right)\right]}{1+\exp\left[1.7a_j\left(\theta_i - b_j\right)\right]} \qquad (5)$$

where a_j is the slope (or item discrimination parameter) of the j-th item, b_j is the location (or item difficulty parameter), c_j is the lower asymptote (or pseudo-guessing parameter), and u_j is the upper asymptote (or carelessness parameter). When the upper asymptote, u_j, is set equal to 1 then the equation forms a 3-PL model.

Although the 4-PL model is considered as the model that best fits the measurement situation, the 4-PL model is not a commonly used IRT model among practitioners and researchers (Kalkan & Cuhadar, 2020; Loken & Rulison, 2010). The reason is due to the difficulty in estimating the upper asymptotes and the unavailability of computer software that practitioners and researchers can use for estimating item and ability parameters under the 4-PL model (Jian et al., 2021; Kalkan & Çuhadar, 2020; Loken & Rulison, 2010). However, recently the 4-PL model has become more popular in the literature on IRT and computerized adaptive testing (CAT), through the development of very powerful computer software programs such as the "mirt" package in R program (Chalmers, 2012). It has been suggested that the use of the 4-PL model provides an advantage in terms of improving the estimation of ability parameters in high-ability test takers under the CAT environment who incorrectly answer early test items (Cheng & Liu, 2015; Culpepper, 2016; Liao et al., 2012, 2012; Loken & Rulison, 2010). Many studies have also contributed to the improvement of the 4-PL IRT model regarding its application and parameter estimation (Culpepper, 2016; Kalkan & Çuhadar, 2020; Liao et al., 2012; Loken & Rulison, 2010; Magis, 2013; Meng et al., 2020; Robitzsch, 2022; Yen et al., 2012).

In this study, our focus is to use the 4-PL IRT model to estimate item parameters. Thus, there are four parameters used to describe item characteristics: item difficulty (b), item discrimination (a), pseudo-guessing (c), and carelessness (u). In this study, an item is said to be easy when $b < -1$, medium when $-1 \leq b < 1$, and hard when $b > 1$ (Georgiev, 2008). Baker and Kim (2017) classify the item discrimination parameter into six categories: very low (0.01 – 0.34), low (0.35 – 0.64), moderate (0.65 – 1.34), high (1.35 – 1.69), very high (> 1.70), and perfect ($+\infty$). In this study, we adapted the classification by Baker and Kim (2017), but we only used three categories, namely poor (< 0.65), fair (0.65 – 1.34), and good (> 1.34). The pseudo-guessing parameter estimates must be low (Barnard-Brak et al., 2018), so that c is less than $1/k$, where k is the number of options, is acceptable (Hambleton et al., 1991; Retnawati, 2014). The carelessness parameter is expected to remain high, so u is not less than 0.90, which is acceptable (Barnard-Brak et al., 2018). This shows that the probability of participants with high abilities answering an item correctly must be relatively high (Barnard-Brak et al., 2018).

Statement of the Problem and Research Questions

A test blueprint or test specification plays an important role in the development of a test and its items. By referring to the test blueprint, test developers or item writers can at least ensure that the test will contain items that can cover all domains, content areas, cognitive processes based on a particular taxonomy (e.g., revised Bloom's taxonomy), and learning objectives that are the focus of the assessment using the test (AERA et al., 2014; Downing, 2006; Idris et al., 2021; Jailani et al., 2023; Omopekunola & Kardanova, 2024). The test blueprint can thus assist the test developer or item writer to enhance content validity (Abdellatif & Al-Shahrani, 2019; Eweda et al., 2020; Kalkbrenner, 2021); in turn, it is expected that content validity evidence can be provided.

Among the components of a test blueprint are specifications regarding the psychometric characteristics of the test and test items and the proportion or distribution of items based on particular aspects (AERA et al., 2014; Downing, 2006). The psychometric characteristic of test items that is generally set in a test blueprint – so-called *a priori* item characteristic – is the level of item difficulty – this is assigned as a category (i.e., easy, medium, and hard) instead of as a numerical value, where this psychometric characteristic reflects how difficult a test is in general as seen from the distribution of test items based on their difficulty level. In test development, it is frequently the case that the test blueprint developer and the item writer are different parties or people, so it is possible that the characteristics of an item referred to in the test blueprint can be perceived differently by the item writer. Even if the test blueprint developer and item writer are the same person, there is no sufficient guarantee that the constructed test items have the characteristics that should be as specified in the test blueprint.

Asking experts in educational measurement and assessment and in the content area or domain relevant to the test to provide their judgments on the alignment of the constructed items with the characteristics that the items should have can be a strategy to address

possible mismatches between the constructed test item and the characteristics that the item should have. This strategy, however, brings its own challenges in relation to the issue of subjectivity (Downing, 2006; Sayin & Bulut, 2024). Item analysis based on empirical data in the form of responses from a number of participants as samples at the trial stage or participants as test targets offers support to the expert judgment strategy. CTT and IRT are two measurement frameworks that can be used simultaneously in item analysis since they can be viewed as complementary measurement frameworks given their respective strengths and weaknesses (see Hambleton et al., 1991; Hambleton & Jones, 1993; Lord, 1980). It is worth noting that although there are a variety of psychometric characteristics of test items (e.g., difficulty and discriminating power), the most relevant *a priori* item characteristic to investigate further in terms of its alignment based on the estimation results under the CTT and IRT frameworks is item difficulty.

The 4-PL model is a dichotomous IRT model that has received more attention in the last two decades, although there remains debate about the interpretation which is considered confusing and application of the model in practical contexts. Barton and Lord's (1981) study began to introduce the 4-PL model through the addition of an upper-asymptote (carelessness) parameter, whose estimated value is less than 1, to the 3-PL model so that it would not severely underestimate the ability estimates of high-ability test takers as a consequence of giving incorrect responses on test items that they should be able to correctly respond to. The results of their study, however, led them to recommend that the use of the 4-PL model is not an urgent matter because it requires a complex computational process in estimating parameters and does not offer a consistent increase in likelihood or a significant change in ability estimates compared to estimation under the 3-PL model. The recommendation provided by Barton and Lord should not be fully adhered to because what they did was more about comparing the goodness-of-fit of models with specified values of u, so the focus was not on estimating the u parameter directly; this has also been noted by Loken and Rulison (2010) and Świst (2015). Given that previous studies have been devoted to exploring the use of the 4-PL model in terms of estimating ability parameters, the current study is directed to further investigate the 4-PL model in relation to estimating item parameters.

In regard to the measurement framework of CTT and the modern test theory associated with Rasch measurement and dichotomous IRT models (i.e., 1-PL, 2-PL, and 3-PL), some studies (e.g., Awopeju & Afolabi, 2016; Fan, 1998; Progar & Sočan, 2008; Setiawati et al., 2023) have empirically demonstrated that the two frameworks are comparable in relation to the item characteristics estimate consisting of item difficulty and discrimination. Lord (1980) has even theoretically proposed mathematical equivalence estimates for item characteristics of item difficulty and discrimination under both measurement frameworks. However, the equivalence estimation of these item characteristics is only valid under specific conditions, including that the distribution of ability parameter estimates follows a normal distribution and there is no identified guessing. It is worth reporting that the approximations Lord (1980) proposed are not for practical use but rather to provide an idea of the nature of item difficulty and discrimination parameters. Despite ample empirical evidence of the comparability of item difficulty and discrimination estimates under CTT and IRT models, the findings of the study by Eleje et al. (2018) suggested the incomparability between the two models. Although MacDonald and Paunonen (2002), through simulations with Monte Carlo technique, have demonstrated that the ability and item difficulty estimates under the two measurement models are comparable, insufficient support for comparability on item discrimination was found. This contradiction in the results of previous studies thus opens up further discussion on the comparability of the two measurement frameworks, especially when the dichotomous IRT model under investigation is the 4-PL model which is considered to be understudied.

The presence of CTT and IRT as measurement frameworks has inevitably offered a variety of applications and advantages in measurement and assessment practices, but it also raises challenges. One of the advantages offered by both is that it makes it possible to investigate the extent to which item characteristics (in terms of difficulty level) align between what is expected by the test developer as set out in the test blueprint and the actual conditions indicated by test taker response data. Since this investigation can reveal the characteristics of the test items, the results of estimating the characteristics of the test items under the CTT and IRT models can be used by this study to further explore the

extent of the comparability of item difficulty and discrimination estimates under the two models. Given the purpose of this study and the gaps that need to be filled, this study thus focused on answering the following research questions (RQs).

RQ1: What are the characteristics of the items used in the Business English test based on the CTT and 4-PL IRT models?

RQ2: How does the perceived item difficulty on the blueprint (*a priori*) align with that when estimated based on the CTT and 4-PL IRT models?

RQ3: How is the comparability of test item difficulty between that estimated with the CTT model and that estimated with the 4-PL IRT model?

RQ4: How is the comparability of test item discrimination between that estimated with the CTT model and that estimated with the 4-PL IRT model?

RQ5: How is the relationship between pseudo-guessing and carelessness parameters estimated under the 4-PL IRT model?

Method

Design and Context of the Study

This quantitative study focused on describing the characteristics of the items from the Business English test and comparing the consistency of these characteristics based on the results of the mapping of the test blueprint (*a priori* item characteristics), CTT model, and 4-PL IRT model. The blueprint of the Business English test was developed by lecturers who teach Business English courses at the Indonesia Open University (Universitas Terbuka, UT) – a public university in Indonesia that facilitates learning through an open and distance learning system. The test blueprint was developed by considering the modules that students learned. The test blueprint that has been developed was then used as a guideline in constructing the test items. The Business English test items were constructed by the blueprint test developers or lecturers from other higher education institutions who teach Business English course or other courses that are highly relevant to the Business English course who were officially appointed to construct the Business English test items. They constructed test items by considering a number of aspects set out in the test blueprint including considering how difficult the test items they were supposed to construct were (see Appendix 1 for *a priori* item characteristics). Item writers generally use their perceptions, knowledge, or experience in constructing the items of the Business English test and adjust the items to the corresponding item difficulty categories as set out in the test blueprint. We unfortunately have no guarantee that all item writers of the Business English test are knowledgeable about the level of item difficulty that can be considered according to the CTT work which is based on the proportion of test takers who can correctly answer an item.

Item characteristics estimated based on CTT were focused on the item difficulty and discrimination statistics. While the item characteristics estimated based on IRT include four parameters: difficulty, discrimination, pseudo-guessing, and carelessness. In the present study, the item difficulty based on *a priori*, CTT, and IRT was classified into three categories: easy, medium, and hard. The item discrimination estimated based on CTT and IRT was also classified into three categories: poor, fair, and good. The pseudo-guessing and carelessness parameters were classified into two categories: acceptable and unacceptable. In addition, the consistency of item characteristics was focused on the item difficulty and discrimination. Specifically for the item difficulty, the difficulty level categories (i.e., easy, medium, and hard) between *a priori*, CTT, and IRT would be compared. As for the item discrimination parameter, we would only compare the parameter estimates based on CTT and IRT.

Participants

This study involved 4,836 students from a public university in Indonesia, the Open University (well known as Universitas Terbuka, UT). Participants were spread across 39 Distance Learning Program Units (*Unit Program Belajar Jarak Jauh*, UPBJJ) organized and managed by Universitas Terbuka. They were students participating in the final examination for the Business English course which be held at the end of 2022. The examination was carried out in each unit using a paper-and-pencil-based test and under strict supervision.

Data Collection

This study used student response data in the final examination of the Business English course. The test blueprint can be seen in Appendix 1. The test consisted

of 30 four-option multiple-choice items with three distractors and one keyed option. The test taker obtained a score of 1 for the correct answer and 0 for the wrong answer. There was no deduction of points when participants answer the wrong item. The test was administered in paper-and-pencil mode. Examinations were strictly supervised to prevent cheating. Student response data was then stored at the Open University Data Center. In this study, we requested permission from the authorities to access student response data in the final examination of Business English. The data we have was the response scores of 4,836 test takers. The data we received was dichotomous data (0 and 1), not raw data. This dichotomous data was the result of scoring the response of the test takers on the Business English final examination. The data consisted of dichotomous scores for each test taker on the 30 existing test items. We conducted a preliminary analysis to check whether all items were feasible for further analysis. In this preliminary analysis, we found three items to have a negative point-biserial correlation coefficient: item 4 ($r_{pbis} = –0.078$), item 16 ($r_{pbis} = –0.020$), and item 26 ($r_{pbis} = –0.134$). In order to maintain stability in item calibration, especially in IRT, we decided to exclude these three items. Finally, in this study we included only 27 items for further analysis.

Data Analysis

In this study, we performed data analysis in several stages. First, we estimated the item statistics using CTT model. At this stage, we obtained the item statistics of difficulty (denoted by p) and discrimination. In this study, the item difficulty index has three categories: easy ($p > 0.7$), medium ($0.3 \leq p \leq 0.7$), and hard ($p < 0.3$). Afterwards, we compared the difficulty level of the mapped items based on the test blueprint that represent *a priori* item characteristics (see Appendix 1) with the difficulty level estimated based on CTT model to determine their alignment. In other words, this way was performed to investigate whether the expected item difficulty level (*a priori*) is consistent with the estimation results based on CTT model. The same procedure was also carried out to compare the *a priori* difficulty level with the estimation results based on IRT under 4-PL model. In this study, the item discrimination parameter also has three categories: poor ($r_{pbis} < 0.1$), fair ($0.1 \leq r_{pbis} \leq 0.3$), and good ($r_{pbis} > 0.3$). The item discrimination category based on CTT model will then be compared with the item discrimination category based on 4-PL IRT model.

Second, we estimated item parameters using IRT. Since the IRT for dichotomous data has several models (i.e., Rasch, 1-PL, 2-PL, 3-PL, and 4-PL), a goodness-of-fit assessment was carried out to determine the best model that fits the data. This assessment includes item-fit and global model-fit analysis. The item-fit analysis was conducted using the signed chi-squared test (Orlando & Thissen, 2000, 2003) where a dichotomous IRT model is said to fit the test takers' response pattern on an item under investigation when the significance value of the statistic for that item is greater than or equal to the significance level used in the current study, which is 0.05. The model that fits the most items indicates that it is the most favorable model to use in estimating the item parameters. We used the results obtained from the item-fit analysis to confirm the results obtained in the global model-fit analysis. We conducted a global model-fit analysis using four fit indices: Akaike information criterion (AIC), Bayesian information criterion (BIC), sample-size-adjusted BIC (SABIC), and root mean square error of approximation (RMSEA). Based on global model-fit analysis, the most favorable model among the other models to be used in estimating item parameters is the one with the smallest estimated value in most of the fit indexes. Table 1 presents the results of the item-fit and global model-fit analyses, where the results of both analyses suggest that the 4-PL model is the model with the most favorable fit given the test taker response patterns on the test items as well as the response patterns on the overall test. Given the results obtained from the goodness-of-fit assessment, the parameters of the Business English test items were thus estimated based on the 4-PL IRT model.

Before the item parameter estimates were executed, we examined whether the assumptions underlying the IRT were satisfied. The three IRT assumptions that we examined were unidimensionality, local independence, and parameter invariance. We reported the results of our investigation on the IRT assumptions' satisfaction at the beginning of the Results section. As the 4-PL model is the model we used to estimate the parameters of the Business English test items, there would be four item parameters that we focus on, namely slope (a), location (b), pseudo-guessing (c), and carelessness (u). The slope parameter represents the item discrimination, and the location parameter represents the item difficulty. In this study, item discrimination parameter which is estimated under the 4-PL IRT

Table 1. The results of item-fit and global model-fit analyses

Model	AIC	SABIC	BIC	RMSEA	Number of fit item
Rasch	155915.9	156018.4	156116.9	0.061246	2
1-PL	155915.9	156018.4	156116.9	0.061236	2
2-PL	152130.1	152328.4	152519.1	0.034318	11
3-PL	151467.8	151765.3	152051.3	0.019266	16
4-PL	**150936.4**	**151333.1**	**151714.4**	**0.018554**	**20**

was categorized into three as adapted from Baker and Kim (2017), namely poor ($a \leq 0.64$), fair ($0.64 < a \leq 1.34$), and good ($a > 1.34$). Item difficulty parameter was also categorized into three, namely easy ($b < -1$), medium ($-1 \leq b \leq 1$), and hard ($b > 1$) (Georgiev, 2008). Pseudo-guessing and carelessness parameters were categorized into acceptable and unacceptable. In this study, pseudo-guessing parameter is acceptable if $c \leq 0.25$ (because the test has four possible answers), while the carelessness parameter is acceptable if $u \geq 0.9$ (Barnard-Brak et al., 2018).

RStudio version 2023.3.1.446 (Posit team, 2023) with three packages was utilized in almost all data analysis in this study, from investigating item characteristics based on the CTT and 4-PL IRT models to examining the satisfaction of IRT assumptions. Three packages we used include 'CTT' package (Willse, 2018) which was used for estimating item statistics based on CTT, 'mirt' package (Chalmers, 2012) which was used for estimating item parameters based on 4-PL IRT model, dan 'psych' package (Revelle, 2023) which was used for examining the satisfaction of the unidimensionality assumption. Some basic commands in R were used within RStudio and combined with 'mirt' package to assess the satisfaction of the parameter invariance assumption. Basic commands were also used to investigate the relationships between carelessness parameter and pseudo-guessing parameter, and the possible relationship of carelessness parameter with difficulty parameter, discrimination parameter, difficulty statistic, and discrimination statistic. Furthermore, some of the graphs presented in this paper were generated using Microsoft Excel.

Results

The current study seeks to investigate the extent of the alignment between test item characteristics in the form of item difficulty expected by the test developer or item writer presented in the test blueprint (so-called *a priori* item characteristics) and the characteristics of these items based on test taker response data. In order to investigate this alignment, test taker response data were analyzed using the CTT and 4-PL models, of which the 4-PL model has been demonstrated to be the best model for estimating item parameters under the IRT framework. Because the item characteristics revealed by the CTT and 4-PL models go beyond item difficulty, this study was thus extended to an exploration of the comparability between item difficulty and discrimination yielded by the two models which in turn reflects the comparability of the two models. Moreover, since the estimation of item characteristics under the 4-PL model yields information on the carelessness parameter, this study further explores the relationship between this parameter and the pseudo-guessing parameter. Since the use of IRT requires the satisfaction of the assumptions underlying IRT, in this section we first present the results of the satisfaction of these assumptions.

Assumptions underlying IRT

First, we reported the unidimensionality assumption of the data. The satisfaction of this assumption was assessed through exploratory factor analysis (EFA) and the scree plot (see Figure 1). Figure 1 shows that there is one factor that has an eigenvalue of more than 1. This indicates that test taker response data supports that there is only one dominant factor measured on the Business English test. Thus, it can be concluded that the unidimensionality assumption is satisfied. Once the unidimensionality assumption is satisfied, the local independence assumption is automatically satisfied (Hambleton et al., 1991). Thus, we believe the test taker response data used in this study supports the satisfaction of local independence assumption.

Assessment of the satisfaction of the parameter invariance assumption is based on the item and ability parameter estimates under the model that best fits the test taker response pattern data. Since the item-fit and

global model-fit analyses indicated the 4-PL model as the most favorable model for estimating item and ability parameters, the test of the assumption of invariance of item parameters involves the slope (*a*), location (*b*), pseudo-guessing (*c*), and carelessness (*u*) parameters. The testing begins by dividing the data into two sets, namely data for test takers in odd order and data for test takers in even order. Each data set was then used to estimate item parameters under the 4-PL model. The item parameters estimated from the first and second data sets under the model are each then presented in scatter plots (see Figures 2(a), 2(b), 2(c), and 2(d)) and we investigated how strongly and significantly related the item parameter estimates from the two data sets are. For the purpose of investigating the satisfaction of ability parameter invariance assumption, we split the data into two sets based on item order, odd-order items and even-order items. Test taker's ability (θ) was then estimated under 4-PL model based on those two data sets. Test takers' abilities estimated from the first data set and the second data set is presented in the scatter plot (see Figure 2(e)) and we investigated the strength and significance of the relationship between the ability parameter estimates based on the two data sets. Figure 2 indicates that the distribution of the estimated item and ability parameters is relatively close to a straight line with a slope of 1 (blue line) and the correlations between the item parameters and ability parameter estimated from the two data sets are fairly strong and significant. These results suggest that there is no invariance problem, both in terms of item parameters and ability parameter. Thus, we have demonstrated the satisfaction of the parameter invariance assumption; this means that the key assumptions underlying IRT have all been satisfied.

Item Difficulty: *A Priori* vs. CTT vs. 4-PL IRT

The first focus of this study was to reveal the extent of alignment between *a priori* item characteristics, particularly in the aspect of item difficulty, and item characteristics estimated based on test taker response data using the CTT and 4-PL IRT models. *A priori* item characteristics in detail are provided in the test blueprint as presented in Appendix 1. Meanwhile, the results of estimating item difficulty based on the CTT and 4-PL models and their categories are presented in detail in Appendix 2. Using the CTT model, our study shows that item difficulty statistic ranged from 0.188 to 0.929 (M = 0.604, SD = 0.207). Meanwhile, using the 4-PL IRT model, our study demonstrates that item difficulty parameter ranged from –1.998 to 2.611 (M = –0.234, SD = 1.106). By referring to the item difficulty category, *a priori* item characteristic suggest that most items (59.26%) have a "medium" difficulty, the CTT model suggest that most items (44.44%) have an "easy" difficulty, and the 4-PL IRT model suggest that most items (55.56%) have a "medium" difficulty.

Figure 3 provides a more detailed distribution of item difficulty across the three categories (i.e., easy, medium, and hard) based on *a priori* item characteristics,

Figure 1. Scree plot of exploratory factor analysis

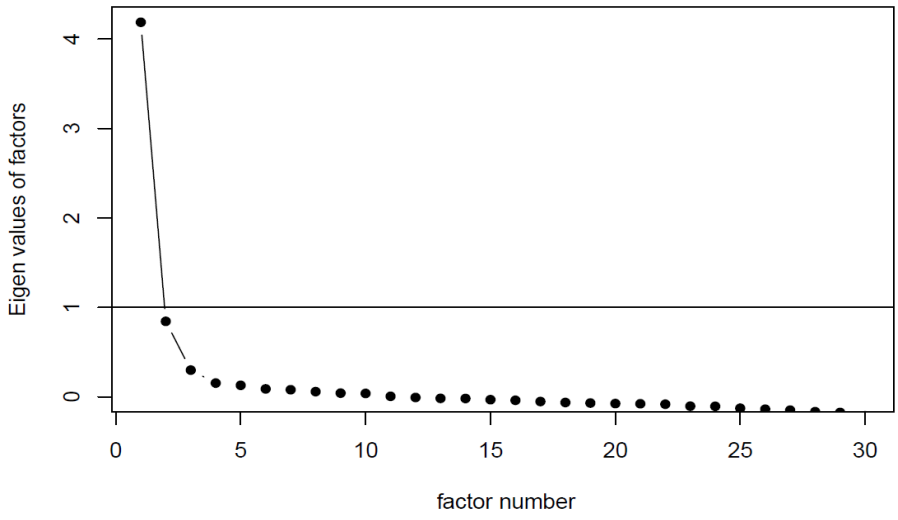
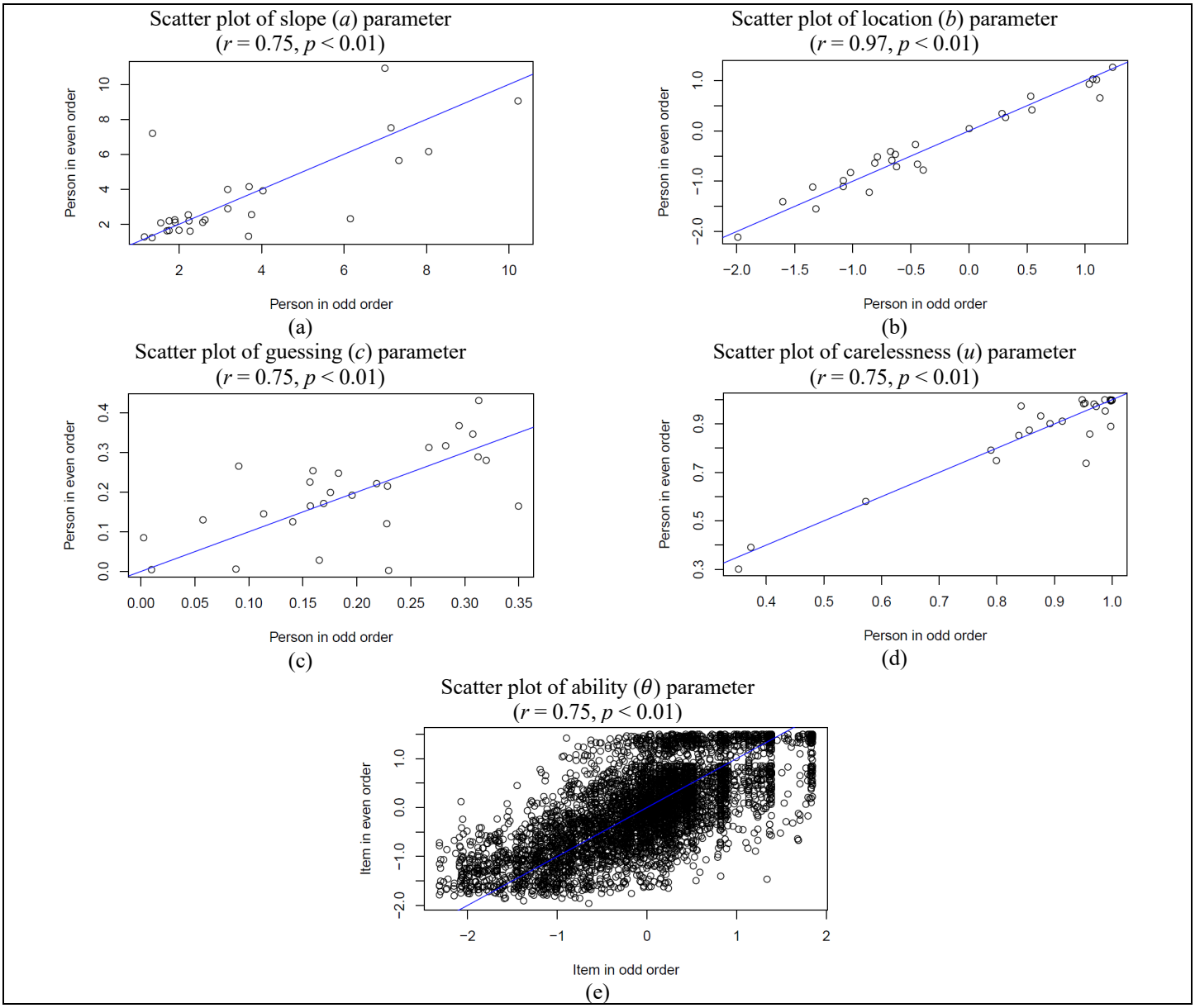


Figure 2. Scatter plots demonstrating parameters invariance: slope (a), location (b), pseudo-guessing (c), carelessness (d), and ability (e)



CTT, and 4-PL IRT. Based on what Figure 3 demonstrates, it is clear that the distribution of item difficulty categories across the three foci of comparison (i.e., *a priori* item characteristics, CTT, and 4-PL IRT) is not consistent. If we investigate further on which model is better in terms of yielding item difficulty estimates that are close to *a priori* item characteristics given the distribution of item difficulty categories, it is clear that the 4-PL IRT model is better than the CTT model.

We present Figure 4 to provide an in-depth look at the distribution of item difficulty based on the three categories and based on the three foci of comparison in this study through the comparison of each item. It was found that there were only six items (22.22%) that had difficulties that fell in the same category between what was expected on the *a priori* item characteristics and those estimated based on the CTT and 4-PL IRT models. The six items are items 9, 14, 18, 21, 22, 28. It was expected by the test developer or item writer that items 9 and 14 have a difficulty in the "hard" category, and this is supported by empirical data estimated based on the CTT and 4-PL IRT models suggesting that both items belong to the "hard" category. The same case was found for the remaining four items, namely 18, 21,

22, and 28, which were confirmed to have difficulties falling in the "medium" category. When the focus is only on the alignment between *a priori* item characteristics and item difficulty estimated based on CTT model, 11 items (40.74%) were identified that aligned with each other in terms of difficulty category. For instance, item 5 with a "medium" difficulty and item 7 with an "easy" difficulty. When it comes to the alignment between the *a priori* item characteristics and the estimated item difficulty under the 4-PL IRT model, we found 9 items (33.33%) to be aligned as indicated by the items falling in the same difficulty category. For instance, items 2 and 3 which both fall into the "medium" difficulty category.

When it comes to the comparison of item difficulty estimates based on the CTT and 4-PL IRT models, there are 17 items (62.96%) that fall in the same difficulty category. For instance, item 1 has an "easy" difficulty, item 6 has a "medium" difficulty, and item 25 has a "hard" difficulty. To further investigate the comparability between the two models in terms of item difficulty estimates, we conducted a correlation analysis with Pearson's correlation on item difficulty estimates under the two models. The results of the analysis revealed that there is a significant negative correlation between the item difficulty estimated under the CTT model and that estimated under the 4-PL IRT model (r = –0.887, p < 0.001). A negative correlation occurs because under the CTT model, the higher the item difficulty on the CTT indicates the easier the item is and the opposite for the case of item difficulty estimates under the IRT model. The linear relationship of the item difficulty estimated under CTT, and 4-PL IRT is presented in Figure 5. The existence of a significant correlation indicates that the estimation results of the item difficulty estimates between CTT and 4-PL IRT are quite consistent and can be substituted for each other. When an item is declared "hard" based on the CTT model, it is likely that the item would fall into the "hard" category based on the estimated item difficulty under the 4-PL IRT model.

Item Discrimination: CTT vs. 4-PL IRT

Given that estimating item characteristics using the CTT model allows us to obtain information on item discrimination, the comparison between the CTT and 4-PL IRT models is extended by taking item discrimination into account. The point-biserial correlation coefficient (r_{pbis}) representing item discrimination under the CTT model and slope parameter (a) representing item discrimination under 4-PL IRT model for each item are presented in detail in Appendix 3. It has been revealed that the item discrimination estimated based on the CTT model ranged from 0.076 to 0.491 (M = 0.332, SD = 0.104), while based on the 4-PL IRT model, it ranged from 1.262 to 6.558 (M = 2.676, SD = 1.485). Based on the CTT model, most of the items (66.67%) have a "good" discrimination and only one item with a "poor" discrimination. Meanwhile, based on the 4-PL IRT model, 88.89% of the items have a

Figure 3. Comparison of item difficulty across three categories based on a priori, CTT, and 4-PL

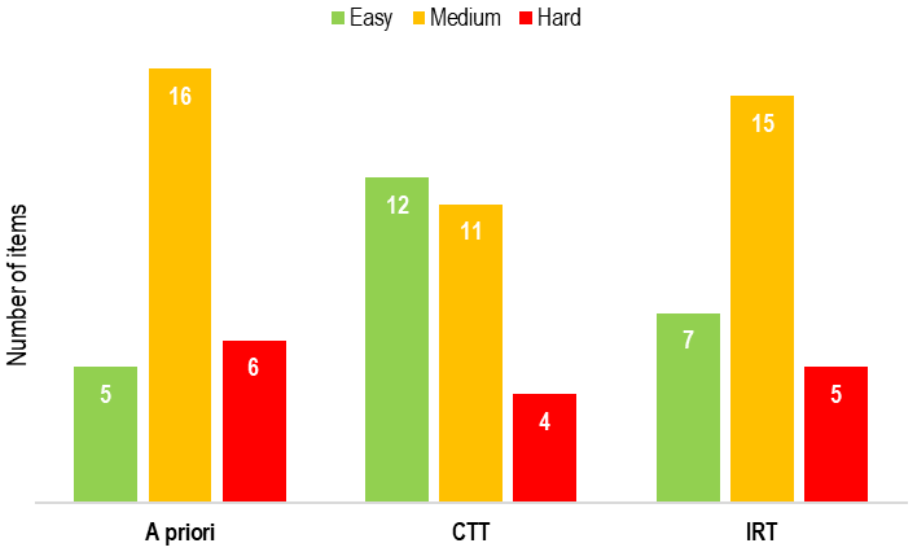


Figure 4. Distribution of item difficulty category based on a priori, CTT, and 4-PL IRT

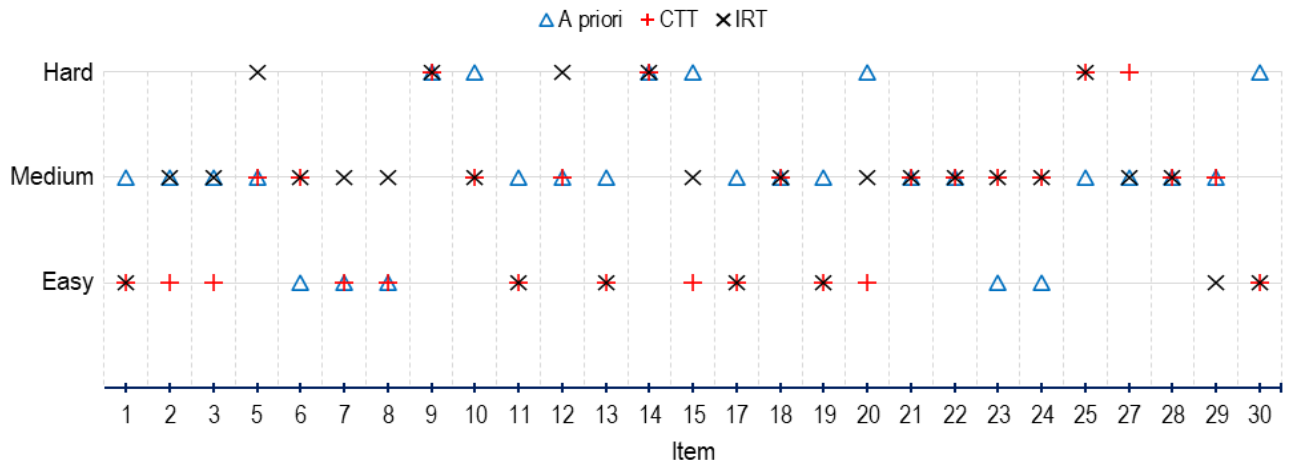
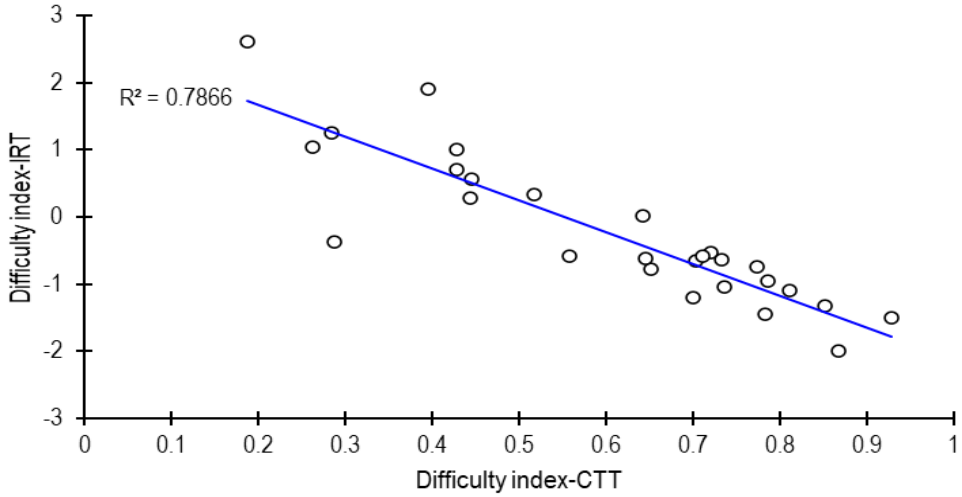


Figure 5. Scatter plot demonstrating correlation between item difficulty based on the CTT and 4-PL IRT models



"good" discrimination and there are no items with a "poor" discrimination.

Figure 6 further presents a comparison of item discrimination estimated under the CTT and 4-PL IRT models based on the distribution of test items across the three discrimination categories. Figure 6 clearly demonstrates that the item discrimination category estimated using the CTT and 4-PL IRT models is inconsistent, but the estimation with the 4-PL IRT model is clearly more favorable because it leads to more items with discrimination falling in the "good" category and the absence of items with discrimination in the "poor" category.

Figure 7 further presents the results regarding the extent of the consistency of an item's discrimination category when the discrimination is estimated based on the CTT model and that when it is estimated based on the 4-PL IRT model. From Figure 7, it can be identified that there are 19 items (70.37%) that consistently fell into the same discrimination category when estimated using both the CTT and 4-PL IRT models. In addition, some items have discrimination that falls into the better discrimination category when the 4-PL IRT model is used to estimate item parameters. For instance, the discrimination of items 5 and 9 fell into the "good" category when item discrimination was estimated using the 4-PL IRT model, whereas when the

discrimination of the two items was estimated using the CTT model, the discrimination of the two items was in the "fair" category. Another example is demonstrated by item 25, where when the CTT model was used to estimate the discrimination of the item, the discrimination of the item fell into the "poor" category. However, when it was estimated using the 4-PL IRT model, the discrimination of the item fell into the "good" category.

In order to gain a better understanding on whether the item discriminations estimated based on the CTT and 4-PL IRT models are interchangeable, a correlation analysis through Pearson's correlation was performed (see Figure 8). The correlation analysis indicates that there is no significant correlation between the item discriminations estimated based on CTT and IRT ($r = -0.175$, $p = 0.383$). This indicates that there is no linear relationship between item discriminations estimated under the CTT and 4-PL IRT models. Thus, the results of item discrimination estimated under the CTT model cannot be used to predict accurately with the least possible error the item discrimination estimated under the 4-PL IRT model. In other words, an item that has discrimination that falls into the "good" category when estimated under the CTT model does not necessarily automatically lead to the discrimination of the item also falling into the "good" category when the 4-PL model is used to estimate item parameters.

Pseudo-guessing and Carelessness Parameters (4-PL IRT)

One of the advantages of using the 4-PL IRT model is that we can estimate the item difficulty and discrimination parameters and also estimate pseudo-guessing and carelessness parameters. In this study, the results of the estimation of pseudo-guessing (c) and carelessness (u) parameters for each item are presented in detail in Appendix 3. The pseudo-guessing parameter estimates ranged from 0.002 to 0.389 ($M = 0.185$, $SD = 0.106$), while the carelessness parameter estimates ranged from 0.384 to 1.000 ($M = 0.917$, $SD = 0.129$).

Figure 9 presents the distribution of pseudo-guessing and carelessness parameters for each item. In this study, items with a pseudo-guessing parameter estimates of less than 0.25 were considered "acceptable". In addition, items with a minimum carelessness estimate of 0.9 are considered "acceptable". Based on Figure 9, 19 items (70.37%) have a pseudo-guessing parameter estimates of less than 0.25 (acceptable), while 18 items (66.67%) are "acceptable" in terms of the carelessness parameter estimates. Item 27 needs more attention because this item has an extreme value of the carelessness parameter ($u = 0.384$) compared to the other items. This indicates that participants with high ability answered this item incorrectly, but on the other hand the probability of correctly guessing this item by participants with low ability is small ($c = 0.127$).

Figure 6. Comparison of item discrimination across there categories between CTT and 4-PL IRT

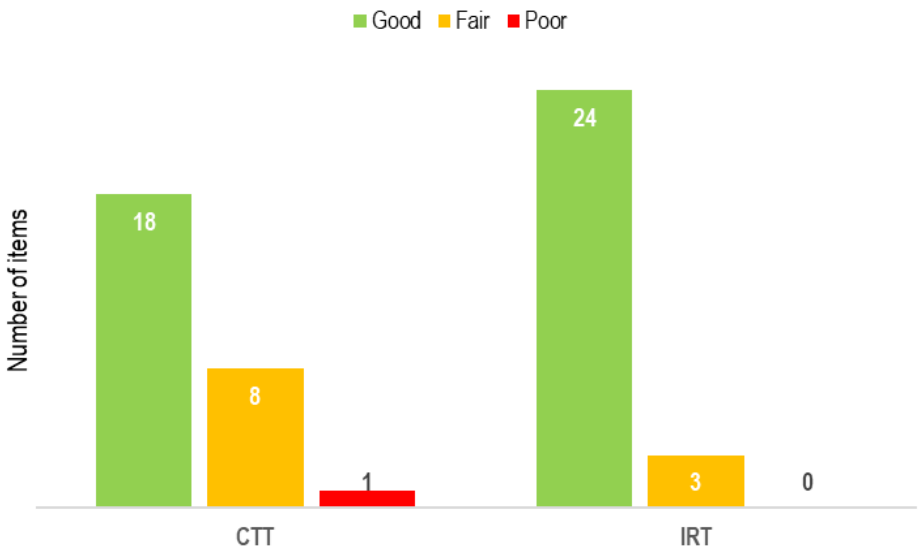


Figure 7. Distribution of item discrimination category based on CTT and 4-PL IRT

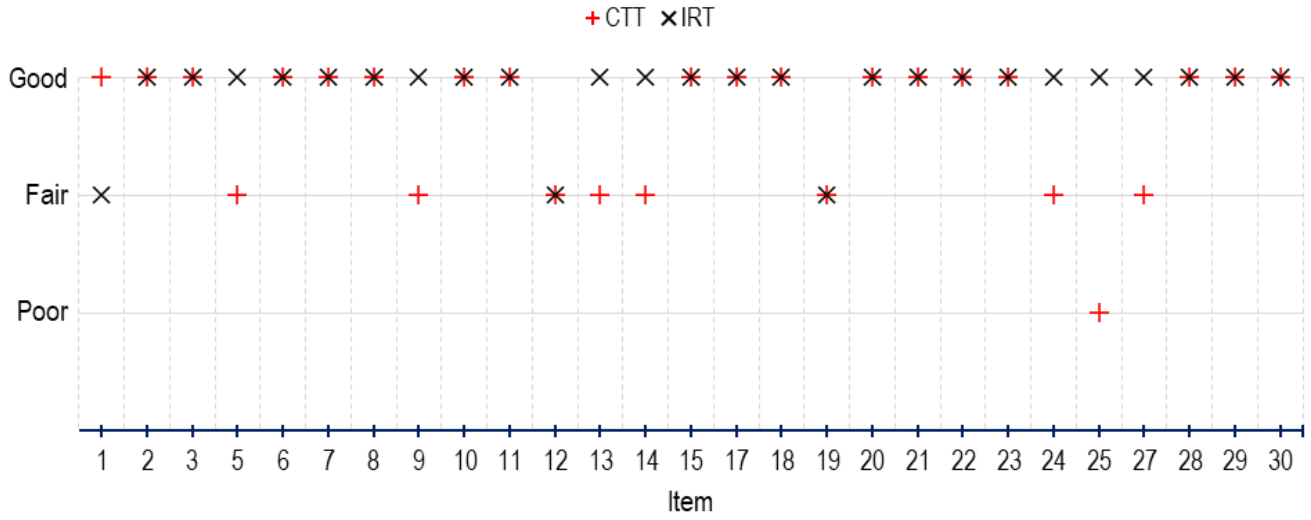


Figure 8. Scatter plot demonstrating correlation between item discrimination based on the CTT and 4-PL IRT models

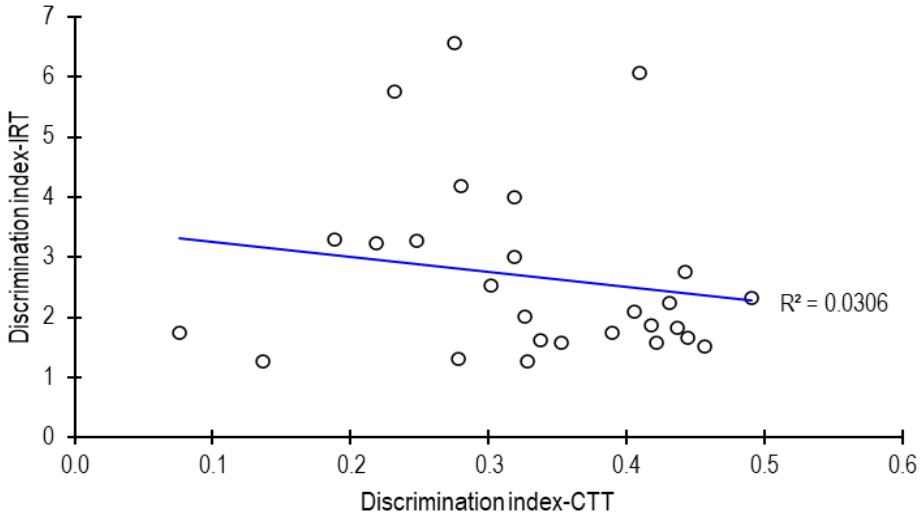


Figure 9 also indicates that there is no relationship pattern between pseudo-guessing and carelessness parameters estimate. Items with "acceptable" pseudo-guessing parameter estimates do not necessarily have "acceptable" carelessness parameter estimates. For example, item 23 is "acceptable" in terms of pseudo-guessing (c = 0.083), but the item is "unacceptable" in terms of carelessness parameter (u = 0.814). In addition, items with "acceptable" carelessness parameter do not necessarily have "acceptable" pseudo-guessing parameter estimates. For example, item 13 is "acceptable" in terms of the carelessness parameter (u = 0.973), but this item is "unacceptable" in terms of the pseudo-guessing parameter (c = 0.389). A detail comparison of "acceptable" and "unacceptable" items between pseudo-guessing and carelessness parameters is presented in Figure 10.

Figure 10 clearly demonstrates that some items (e.g., items 14 and 23) are "acceptable" for the pseudo-guessing parameter estimates but "unacceptable" for the carelessness parameter estimates. On the other hand, some items (e.g., items 5 and 10) are "acceptable" for the carelessness parameter estimates but "unacceptable" for the pseudo-guessing parameter estimates. Therefore, based on these findings, the results of estimating pseudo-guessing and carelessness para-

meters from the test are not as expected. We then performed a correlation analysis through Pearson's correlation to examine whether there is a relationship between the pseudo-guessing and carelessness parameters estimate. The analysis suggested that there was no significant correlation between pseudo-guessing and carelessness parameters estimate (r = 0.056, p = 0.783). This result indicates clearly that there is no linear relationship between pseudo-guessing and carelessness parameters estimate (Figure 11).

Although it is not part of the main research question in this study, it is also worth investigating which item statistics or parameters other than the pseudo-guessing parameter might show a strong and significant relationship with the carelessness parameter. By using correlation analysis through Pearson's correlation, we found that there was insufficient evidence to suggest that item difficulty statistic (r = 0.317, p = 0.107), item discrimination statistic (r = 0.328, p = 0.094), item difficulty parameter (r = 0.054, p = 0.789), and item discrimination parameter (r = –0.150, p = 0.455) have a strong and significant relationship with the carelessness parameter.

Figure 9. Distribution of pseudo-guessing and carelessness parameter estimates for each item

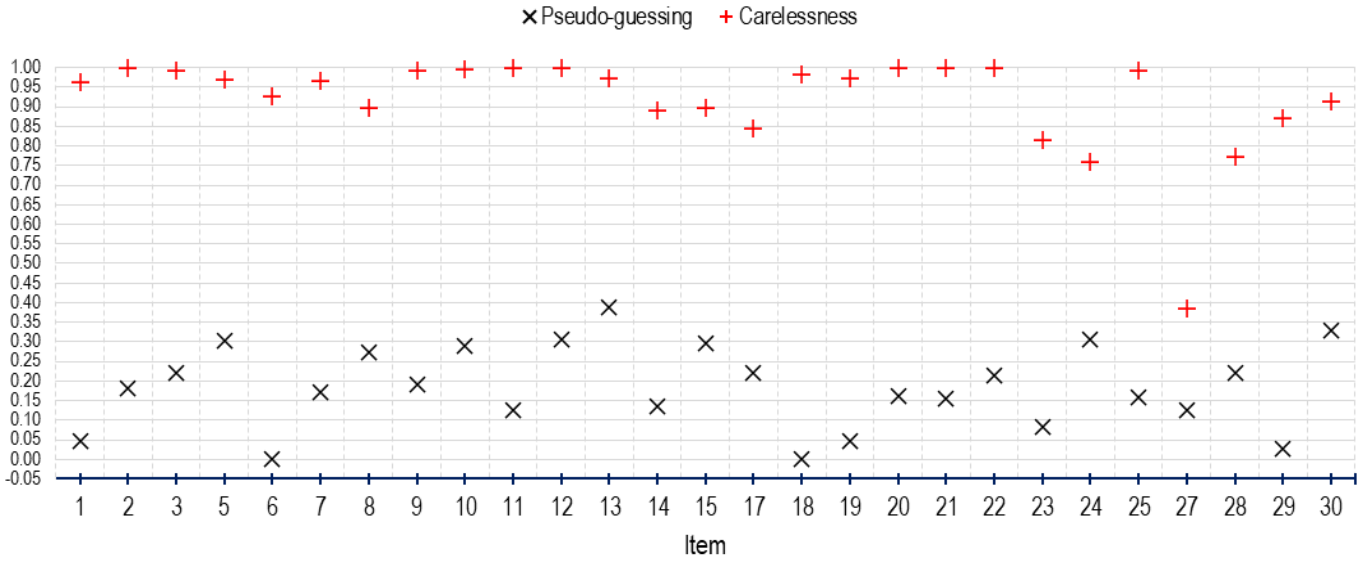


Figure 10. Distribution of pseudo-guessing and carelessness categories for each item

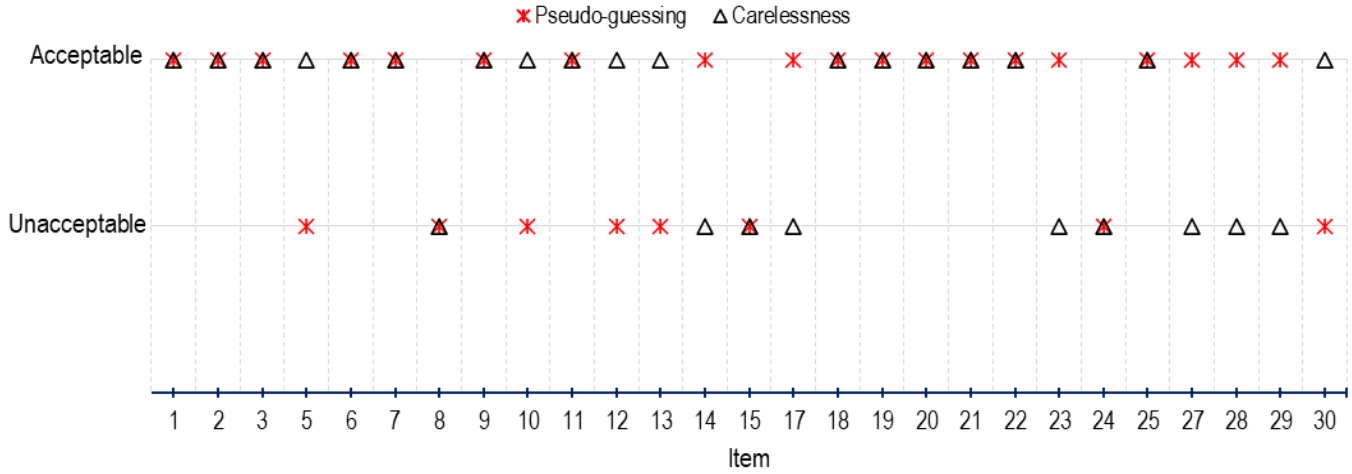
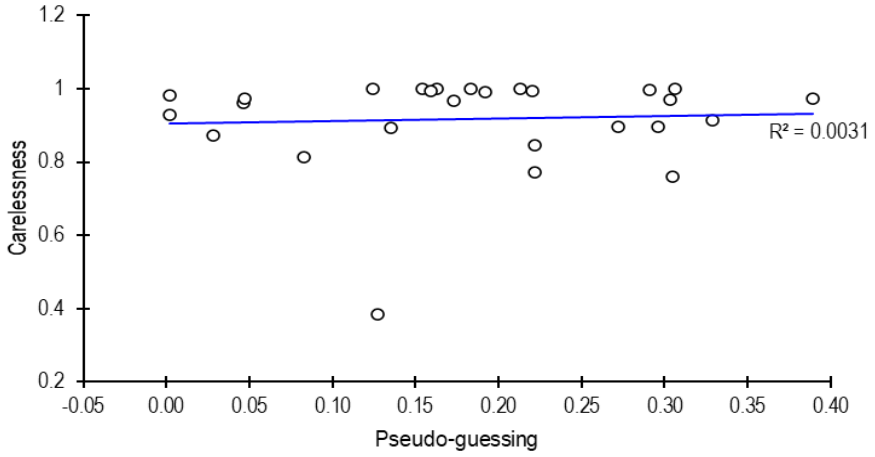


Figure 11. Scatter plot demonstrating correlation between pseudo-guessing and carelessness parameters estimate



Discussion

A test blueprint is essential in test development because it serves as a guide for item writers to construct test items including aligning the characteristics of constructed test items with the specifications of the test and test items. One of the specifications is item characteristics in the form of item difficulty level in the category (*a priori* item characteristic) and the distribution of items in the test based on difficulty level which represents the difficulty level of the test. Since the item-by-item interpretation of easy, medium, or hard by item writers can be influenced by their perceptions, knowledge, and experience, support based on empirical data through analysis using the educational measurement framework (i.e., CTT and IRT) is needed to ensure that test items and tests actually have the characteristics they are supposed to have.

Our study seeks to unveil the extent to which the *a priori* item characteristics of the Business English test align with the item difficulty estimated using the CTT and 4-PL IRT models based on empirical data. It is widely understood that in most tests, especially norm-referenced achievement tests, the use of medium difficulty items in the highest proportion (dominant) is highly recommended (Allen & Yen, 1979; Hambleton & Swaminathan, 1985; Quaigrain & Arhin, 2017; Reynolds et al., 2009). Accordingly, it is reasonable that the *a priori* item characteristics of the Business English test place items with medium difficulty as the dominant ones. Analysis of student response data on the test leads to inconsistent results. Analysis using the CTT model suggests the dominance of easy items although in number there is only one item difference with the medium category, while analysis using the 4-PL IRT model shows the dominance of items with a medium difficulty level. This finding indicates that in terms of item dominance based on difficulty level, the *a priori* item characteristic on the Business English test is supported by empirical data when the analysis used is the 4-PL IRT model.

When it comes to the alignment between *a priori* item characteristics and empirical data based on the distribution of test items in terms of the three difficulty categories, the 4-PL model can provide more support for the alignment. However, when the alignment is based on the difficulty category of each test item, our results show that the estimation of item difficulty with the CTT model provides slightly better support for the alignment than the 4-PL model. Given that less than 50% of the test items showed alignment of item difficulty category between *a priori* with the CTT model, *a priori* with the 4-PL IRT model, and *a priori* with both the CTT and 4-PL IRT models, it is not worth asserting that the characteristics of the items used in the Business English test align with the characteristics expected by the test developer. The results of the present study provide additional support to previous studies (e.g., Sayin & Bulut, 2024) suggesting that it is not easy to construct test items that have the same characteristics as those expected in the test blueprint and those estimated based on empirical data using the CTT and IRT models. Post examination item analysis using the CTT and IRT models performed in the present study provides support for alternative efforts to improve the accuracy of revealing the item characteristics, especially

item difficulty. The results of such item analysis can also be complementary to the results obtained from judgments that subject matter experts provide through comparative and non-comparative methods related to the item difficulty level that has been demonstrated to enhance the alignment between perceived and estimated item difficulty in a real or future test (Attali et al., 2014; Berenbon & McHugh, 2023; Swaminathan et al., 2003).

In the current study, we further explored the extent of comparability between CTT and IRT when a 4-PL model is used in terms of item difficulty and discrimination estimates. In terms of item difficulty estimates, this study has revealed a strong and significant correlation between item difficulty estimates estimated based on the CTT model and those estimated based on the 4-PL IRT model, indicating that item difficulty estimated using the CTT and IRT models are comparable. This finding is consistent with the findings of previous studies (Adegoke, 2013; Awopeju & Afolabi, 2016; Bichi et al., 2019; Progar & Sočan, 2008; Setiawati et al., 2023) but different from the findings of other studies (Ayanwale et al., 2018; Eleje et al., 2018).

Given that the estimation of item difficulty under the CTT model is comparable to that under the 4-PL IRT model, the results of item difficulty estimation based on the two models and its category can be used by test developers or item writers to develop items similar to a particular item for future test so that it is expected that the estimation and category of difficulty of the item are similar to those of the reference items. The comparability in item difficulty based on the two measurement models demonstrated by our study also leads to the suggestion of providing training to test item writers that enables them to understand predicting the difficulty of an item based on the estimated proportion of test takers who will be able to answer the item correctly. This understanding, which in essence represents the concept of item difficulty under the CTT model, has also been shown by Swaminathan et al. (2003) could lead to improved accuracy of item difficulty estimation under the IRT model.

While item difficulty estimates lead to the result that CTT and IRT are comparable, inconsistent estimates of item discrimination using CTT and 4-PL IRT have been identified in our study. This study also found that there was no significant correlation between the item discrimination estimates obtained from the use of the CTT model and those from the use of the 4-PL IRT model. This finding is consistent with previous studies (Ayanwale et al., 2018; Eleje et al., 2018) which found that the item discrimination estimated based on the CTT and IRT models are not comparable. However, this finding is different from the findings of previous studies (i.e., Adegoke, 2013; Awopeju & Afolabi, 2016; Bichi et al., 2019; Progar & Sočan, 2008; Setiawati et al., 2023). Our findings confirm that an item which falls in the "good" category for discrimination based on the CTT model may not necessarily fall in the same category based on the 4-PL IRT model, and vice versa. However, this study revealed that estimation of item discrimination under the IRT model leads to more items that have discrimination in the "good" category, indicating that the use of the 4-PL IRT model is more favorable than the use of the CTT model to estimate item characteristics, in particular item discrimination. The insufficient support for the comparability of the CTT and 4-PL IRT models based on item discrimination raises greater challenges for further exploration of the mathematical equivalence estimates of CTT and 4-PL IRT item discrimination as Lord (1980) has undertaken. This challenge paves an avenue for future studies to provide more evidence for the comparability between CTT and IRT, especially on the 4-PL IRT model, including exploring the possibility of item characteristic equivalence for the two models when there is sufficient evidence that the two models are comparable.

The preliminary analysis that led to the 4-PL model being selected as the most favorable for estimating item parameters in this study presents both opportunities and challenges. The 4-PL model allows us to justify the quality of a test item in more depth as more characteristics of the item are captured. It has been suggested that when it comes to a multiple-choice test, the quality of the test item should not only be justified based on difficulty and discrimination but also on the effects of (pseudo-)guessing and carelessness; and the 4-PL model accommodates for this purpose (Barnard-Brak et al., 2018; Kalkan & Çuhadar, 2020; Liao et al., 2012). Items with a high probability of guessing benefit test takers with low ability even though they actually do not have sufficient knowledge to answer them correctly (Barnard-Brak et al., 2018; Liao et al., 2012). Meanwhile, items with low carelessness parameter estimates have the potential to punish test takers with high abilities even though they should have been able to answer

these items correctly (Barnard-Brak et al., 2018; Kalkan & Çuhadar, 2020; Liao et al., 2012). Items with poor pseudo-guessing and carelessness parameters have the potential to disadvantage test takers when they are used in the measurement (Liao et al., 2012).

In this study, we also investigated whether there is a linear relationship between pseudo-guessing and carelessness parameters. Our study has demonstrated that there is no significant correlation between these two parameters. This indicates that even though an item has acceptable pseudo-guessing parameter, it does not necessarily have acceptable carelessness parameter and vice versa. In other words, we can say that the pseudo-guessing and carelessness parameters are independent. This finding is in line with the study of Antoniou et al. (2022). However, these parameters are equally important to ensure the quality of the items while at the same time giving participants a sense of fairness when used in the measurement (Barnard-Brak et al., 2018; Kalkan & Çuhadar, 2020; Liao et al., 2012).

The results of this study can be used as a guide for test developers in developing tests for educational measurement and assessment. Although item analysis using the CTT model is considered sufficient, the 4-PL IRT model seems more powerful and precise in describing item quality. Item analysis using 4-PL IRT model can also improve the accuracy of measurement results because it takes into account the effects of guessing – due to anomalous behavior shown by students with low abilities which is demonstrated by their success in answering difficult test items correctly – and carelessness – due to the anomalous behavior shown by high-ability students which is indicated by their failure to answer easy test items – of test takers, where these two things cannot be addressed with the CTT model. We believe that this study contributes to the literature regarding the application of CTT and IRT in test development. In addition, through this study we hope that the use of the 4-PL IRT model in educational measurement and assessment would become increasingly popular, given that it can more accurately provide information about test takers' abilities because it accommodates anomalous behavior of test takers in responding to test items as has been suggested by previous studies (Doğruöz & Arikan, 2020; Liao et al., 2012; Waller & Feuerstahler, 2017).

It would also be worthwhile to further discuss the challenges presented by the use of the 4-PL IRT model to estimate the parameters of an item. The challenges are more about the interpretation of the carelessness parameter estimate of an item, especially when it exhibits such a low value, and the implications for the development of a test and test items. If we revisit the carelessness parameter estimate of item 27, it is difficult to provide a reasonable and satisfying interpretation on the estimate of around 0.384 – is considerably low when compared to the carelessness parameter estimation values of the other items which vary from 0.761 to 1.00 – when it is associated with the nature of the parameter demonstrating the carelessness that high-ability test takers exhibit in responding to the item. This difficulty is coupled with the results of the empirical data used in this study indicating insufficient statistical support for the relationship between the carelessness parameter and other item statistics or parameters and no anomalies were detected in the parameter estimates or statistics of item 27 other than in the carelessness parameter estimates. It is evident that when the estimated carelessness parameter is still associated with person-specific, that is carelessness committed by high-ability test takers due to several possible factors (e.g., test medium or delivery format and lack of adequate time to take the test) so that they answer incorrectly an item would provide less valuable information for test developers or item writers in improving the quality of items to be used in a test. This challenge could possibly be one of the barriers to the use of the 4-PL model in operational work concerning the provision of assistance in the development of quality test and test items.

A number of studies have also found that some test items have very low carelessness parameter estimates (e.g., Barnard-Brak et al., 2018; Doğruöz & Arikan, 2020; Pardede et al., 2023). Unfortunately, it is hard to obtain further explanations from these existing studies on the interpretation and implications of such low carelessness parameter estimates of a particular item other than the explanation of the carelessness of high-ability test takers that leads them to provide wrong answers to test items that should be answered correctly. An explanation for the possible reasons that lead high-ability test takers to be extremely careless in responding to an item as indicated by the low carelessness parameter estimate is also difficult to obtain. The only study we found that provides an explanation for the relationship between the estimation of an item's carelessness parameter and signals of item-writing flaws is

that conducted by Świst (2015). She demonstrated that a carelessness parameter estimate of an item that is very low relative to other items can signal that the item is defective in terms of its format (i.e., general item-writing, stem construction, and option development). However, from this, he recognized that to identify whether an item is flawed it is sufficient to use a simpler dichotomous IRT model that is easier to interpret and also through a qualitative analysis by considering the guidelines or format in constructing a good multiple-choice item suggested by Rodriguez (1997).

This study has limitations to report that the number of items and samples we used to estimate the item statistics or parameters was fixed. Consequently, this condition made it impossible for us to investigate the effect of the number of items on the comparison of the estimation results using the CTT and 4-PL IRT models. Likewise with sample size, the effect of sample size on the estimation results of item statistics or parameters also could not be further investigated in this study. We hope that these two crucial issues will be of concern to future researchers. Future studies are expected to conduct simulation studies to investigate the two issues we have mentioned. Furthermore, it is also important to conduct empirical studies in other fields to enrich the literature regarding the comparison of item statistics or parameters estimation results using the CTT and 4-PL IRT models. Above all, this study has one more limitation in that it could not further investigate the quality of the distractors of item 27 that might reveal the contribution of the carelessness parameter to the test item quality because the data we could have and access was only in dichotomous form. Therefore, future studies are expected to pay attention to the quality of distractors in exploring item characteristics based on the 4-PL IRT model, their interpretation, and implications in the development of a test and test items. Since the quality of the distractors could not be explored through this study, this study was limited to further demonstrating the characteristics of the items through providing more item parameters that could lead to more accurate estimates of the ability of test takers.

Conclusions

This study found that the results of item difficulty estimation using the CTT and 4-PL IRT models were comparable, but the *a priori* difficulty level categories and the estimation results using the CTT and 4-PL IRT were inconsistent. These results confirm that what is expected by the test developer regarding the level of item difficulty is not in accordance with the estimation results using empirical data. The results of estimation of item discrimination using the CTT and 4-PL IRT models were not comparable, but estimation under the 4-PL IRT model yielded more items with good discrimination categories so that the model is considered more favorable than the CTT model. In addition, the pseudo-guessing and carelessness parameters estimated using the 4-PL IRT model are independent. These two parameters need to be considered to ensure that the test items accurately measure the actual abilities of test takers. Although the CTT and simpler dichotomous IRT models are deemed sufficient for providing information regarding item quality, we argue that when the 4-PL IRT model is the best fit to the data, the use of the 4-PL model may provide the potential to reveal item characteristics in a deeper way, although more study is needed on this matter, which could lead to an improvement in the accuracy of test taker ability estimates. The CTT and IRT frameworks, especially the 4-PL IRT model, can be utilized to complement each other in addition to qualitative analysis in the test development processes.

References

Abdellatif, H., & Al-Shahrani, A. M. (2019). Effect of blueprinting methods on test difficulty, discrimination, and reliability indices: Cross-sectional study in an integrated learning program. *Advances in Medical Education and Practice*, *10*, 23–30. https://doi.org/10.2147/AMEP.S190827

Adegoke, B. A. (2013). Comparison of item statistics of physics achievement test using classical test and item response theory frameworks. *Journal of Education and Practice*, *4*(22), 87–96. https://www.iiste.org/Journals/index.php/JEP/article/view/8331

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association (AERA).

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Wadsworth.

Antoniou, F., Alkhadim, G., Mouzaki, A., & Simos, P. (2022). A psychometric analysis of Raven's colored progressive matrices: Evaluating guessing

and carelessness using the 4PL item response theory model. *Journal of Intelligence*, *10*(1), 1–14. https://doi.org/10.3390/jintelligence10010006

Attali, Y., Saldivia, L., Jackson, C., Schuppan, F., & Wanamaker, W. (2014). Estimating item difficulty with comparative judgments. *ETS Research Report Series*, *2014*(2), 1–8. https://doi.org/10.1002/ets2.12042

Awopeju, O. A., & Afolabi, E. R. I. (2016). Comparative analysis of classical test theory and item response theory-based item parameter estimates of senior school certificate mathematics examination. *European Scientific Journal*, *12*(28), 263–284. https://doi.org/10.19044/esj.2016.v12n28p263

Ayanwale, M. A., Adeleke, J. O., & Mamadelo, T. I. (2018). An assessment of item statistics estimates of basic education certificate examination through classical test theory and item response theory approach. *International Journal of Educational Research Review*, *3*(3), 93–105. https://doi.org/10.24331/ijere.452555

Baker, F. B. (2001). *The basics of item response theory* (2nd ed). ERIC Clearinghouse on Assessment and Evaluation.

Baker, F. B., & Kim, S.-H. (2017). *The basics of item response theory using R*. Springer International Publishing. https://doi.org/10.1007/978-3-319-54205-8

Barnard-Brak, L., Lan, W. Y., & Yang, Z. (2018). Differences in mathematics achievement according to opportunity to learn: A 4PL item response theory examination. *Studies in Educational Evaluation*, *56*, 1–7. https://doi.org/10.1016/j.stueduc.2017.11.002

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

Berenbon, R. F., & McHugh, B. C. (2023). Do subject matter experts' judgments of multiple-choice format suitability predict item quality? *Educational Measurement: Issues and Practice*, *42*(3), 13–21. https://doi.org/10.1111/emip.12570

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*(5c), 1260–1266. https://doi.org/10.35940/ijeat.E1179.0585C19

Birnbaum, A. (2008). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 396–424). Information Age Publishing.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Cheng, Y., & Liu, C. (2015). The effect of upper and lower asymptotes of IRT models on computerized adaptive testing. *Applied Psychological Measurement*, *39*(7), 551–565. https://doi.org/10.1177/0146621615585850

Cohen, R. J., & Swerdlik, M. E. (2018). *Psychological testing and assessment: An introduction to tests and measurement* (9th ed.). McGraw-Hill Education.

Crocker, L. M., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

Culpepper, S. A. (2016). Revisiting the 4-parameter item response model: Bayesian estimation and application. *Psychometrika*, *81*(4), 1142–1163. https://doi.org/10.1007/s11336-015-9477-6

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. CRC Press.

Doğruöz, E., & Arikan, Ç. A. (2020). Comparison of different ability estimation methods based on 3 and 4PL item response theory. *Pamukkale University Journal of Education*, *50*(1), 50–69. https://doi.org/10.9779/pauefd.585774

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3–25). Lawrence Elbaum Associates.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice-Hall.

Eleje, L. I., Onah, F. E., & Abanobi, C. C. (2018). Comparative study of classical test theory and item response theory using diagnostic quantitative economics skill test item analysis results. *European Journal of Educational & Social Sciences*, *3*(1), 57–75. https://dergipark.org.tr/en/pub/ejees/issue/40156/477675

Eweda, G., Bukhary, Z. A., & Hamed, O. (2020). Quality assurance of test blueprinting. *Journal of*

Professional Nursing, *36*(3), 166–170. https://doi.org/10.1016/j.profnurs.2019.09.001

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*(3), 357–381. https://doi.org/10.1177/0013164498058003001

Georgiev, N. (2008). Item analysis of C, D and E series from Raven's standard progressive matrices with item response theory two-parameter logistic model. *Europe's Journal of Psychology*, *4*(3), 1–15. https://doi.org/10.5964/ejop.v4i3.431

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

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Springer Science+Business Media.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation* (8th ed.). Allyn & Bacon.

Hu, Z., Lin, L., Wang, Y., & Li, J. (2021). The integration of classical testing theory and item response theory. *Psychology*, *12*(1), 1397–1409. https://doi.org/10.4236/psych.2021.129088

Idris, T., Ferazona, S., & Safitri, H. (2021). Profile of the ability of prospective Biology teachers in making question instruments using Bloom's taxonomy. *REID (Research and Evaluation in Education)*, *7*(2), 177–185. https://doi.org/10.21831/reid.v7i2.44903

Jailani, J., Retnawati, H., Rafi, I., Mahmudi, A., Arliani, E., Zulnaidi, H., Hamid, H. S. A., & Prayitno, H. J. (2023). A phenomenological study of challenges that prospective mathematics teachers face in developing mathematical problems that require higher-order thinking skills. *Eurasia Journal of Mathematics, Science and Technology Education*, *19*(10), 1–21. https://doi.org/10.29333/ejmste/13631

Jian, X., Buyun, D., & Yuanping, D. (2021). The robust estimation of examinee ability based on the four-parameter logistic model when guessing and carelessness responses exist. *PLOS ONE*, *16*(4), e0250268. https://doi.org/10.1371/journal.pone.0250268

Kalkan, Ö. K., & Çuhadar, İ. (2020). An evaluation of 4PL IRT and DINA models for estimating pseudo-guessing and slipping parameters. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi* [*Journal of Measurement and Evaluation in Education and Psychology*], *11*(2), 131–146. https://doi.org/10.21031/epod.660273

Kalkbrenner, M. T. (2021). A practical guide to instrument development and score validation in the social sciences: The MEASURE approach. *Practical Assessment, Research & Evaluation*, *26*(1), 1–18. https://doi.org/10.7275/svg4-e671

Kartowagiran, B., Mardapi, D., Purnama, D. N., & Kriswantoro, K. (2019). Parallel tests viewed from the arrangement of item numbers and alternative answers. *REID (Research and Evaluation in Education)*, *5*(2), 169–182. http://doi.org/10.21831/reid.v5i2.23721

Lahza, H., Smith, T. G., & Khosravi, H. (2023). Beyond item analysis: Connecting student behaviour and performance using e-assessment logs. *British Journal of Educational Technology*, *54*(1), 335–354. https://doi.org/10.1111/bjet.13270

Liao, W.-W., Ho, R.-G., Yen, Y.-C., & Cheng, H.-C. (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

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

MacDonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, *62*(6), 921–943. https://doi.org/10.1177/0013164402238082

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

Magno, C. (2009). Demonstrating the difference between classical test theory and item response

theory using derived test data. *International Journal of Educational and Psychological Assessment*, *1*(1), 1–11. https://files.eric.ed.gov/fulltext/ED506058.pdf

McMillan, J. H. (2000). *Fundamental assessment principles for teachers and school administrators*. *7*(8), 1–5. https://doi.org/10.7275/5kc4-jy05

Meng, X., Xu, G., Zhang, J., & Tao, J. (2020). Marginalized maximum a posteriori estimation for the four-parameter logistic model under a mixture modelling framework. *British Journal of Mathematical and Statistical Psychology*, *73*(S1), 51–82. https://doi.org/10.1111/bmsp.12185

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching* (10th ed.). Pearson Education.

Mohajan, H. K. (2017). Two criteria for good measurements in research: Validity and reliability. *Annals of Spiru Haret University*, *17*(4), 59–82. https://doi.org/10.26458/1746

Odukoya, J. A., Adekeye, O., Igbinoba, A. O., & Afolabi, A. (2018). Item analysis of university-wide multiple choice objective examinations: The experience of a Nigerian private university. *Quality & Quantity*, *52*(3), 983–997. https://doi.org/10.1007/s11135-017-0499-2

Omopekunola, M. O., & Kardanova, E. Y. (2024). Automatic generation of physics items with large language models (LLMs). *REID (Research and Evaluation in Education)*, *10*(2), 168–185. https://doi.org/10.21831/reid.v10i2.76864

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64. https://doi.org/10.1177/01466216000241003

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S - X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*(4), 289–298. https://doi.org/10.1177/0146621603027004004

Pardede, T., Santoso, A., Diki, D., Retnawati, H., Rafi, I., Apino, E., & Rosyada, M. N. (2023). Gaining a deeper understanding of the meaning of the carelessness parameter in the 4PL IRT model and strategies for estimating it. *REID (Research and Evaluation in Education)*, *9*(1), 86–117. https://doi.org/10.21831/reid.v9i1.63230

Posit team. (2023). *RStudio: Integrated development environment for R* (2023.3.1.446). Posit Software, PBC.

Primi, R. (2018). Using four-parameter item response theory to model human figure drawings. *Avaliação Psicológica*, *17*(4), 473–483. https://doi.org/10.15689/ap.2018.1704.7.07

Progar, Š., & Sočan, G. (2008). An empirical comparison of item response theory and classical test theory. *Psihološka Obzorja/Horizons of Psychology*, *17*(3), 5–24. http://psiholoska-obzorja.si/arhiv_clanki/2008_3/progar_socan.pdf

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, *4*(1), 1–11. https://doi.org/10.1080/2331186X.2017.1301013

Rafi, I., Retnawati, H., Apino, E., Hadiana, D., Lydiati, I., & Rosyada, M. N. (2023). What might be frequently overlooked is actually still beneficial: Learning from post national-standardized school examination. *Pedagogical Research*, *8*(1), em0145. https://doi.org/10.29333/pr/12657

Retnawati, H. (2014). *Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana* [*Item response theory and its applications: For researchers, measurement and testing practitioners, graduate students*]. Parama Publishing.

Revelle, W. (2023). *psych: Procedures for psychological, psychometric, and personality research* (R package version 2.3.3) [Computer software]. https://cran.r-project.org/package=psych

Reynolds, C. R., Livingston, R. B., & Willson, V. (2009). *Measurement and assessment in education* (2nd ed.). Pearson Education.

Robitzsch, A. (2022). Four-parameter guessing model and related item response models. *Mathematical and Computational Applications*, *27*(6), 95. https://doi.org/10.3390/mca27060095

Rodriguez, M. C. (1997). *The art & science of item-writing: A meta-analysis of multiple-choice item format effects*. The Annual Meeting of the American Educational Research Association, Chicago, IL.

Santoso, A., Pardede, T., Apino, E., Djidu, H., Rafi, I., Rosyada, M. N., Retnawati, H., & Kassymova, G. K. (2022). Polytomous scoring correction and its effect on the model fit: A case of item response theory analysis utilizing R. *Psychology, Evaluation,*

and Technology in Educational Research, *5*(1), 1–13. https://doi.org/10.33292/petier.v5i1.148

Santoso, A., Pardede, T., Djidu, H., Apino, E., Rafi, I., Rosyada, M. N., & Hamid, H. S. A. (2022). The effect of scoring correction and model fit on the estimation of ability parameter and person fit on polytomous item response theory. *REID (Research and Evaluation in Education)*, *8*(2), 140–151. https://doi.org/10.21831/reid.v8i2.54429

Sayin, A., & Bulut, O. (2024). The difference between estimated and perceived item difficulty: An empirical study. *International Journal of Assessment Tools in Education*, *11*(2), 368–387. https://doi.org/10.21449/ijate.1376160

Setiawati, F. A., Amelia, R. N., Sumintono, B., & Purwanta, E. (2023). Study item parameters of classical and modern theory of differential aptitude test: Is it comparable? *European Journal of Educational Research*, *12*(2), 1097–1107. https://doi.org/10.12973/eu-jer.12.2.1097

Subali, B., Kumaidi, K., & Aminah, N. S. (2021). The comparison of item test characteristics viewed from classic and modern test theory. *International Journal of Instruction*, *14*(1), 647–660. https://doi.org/10.29333/iji.2021.14139a

Swaminathan, H., Hambleton, R. K., Sireci, S. G., Xing, D., & Rizavi, S. M. (2003). Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates. *Applied Psychological Measurement*, *27*(1), 27–51. https://doi.org/10.1177/0146621602239475

Świst, K. (2015). Item analysis and evaluation using a four-parameter logistic model. *Edukacja*, *3*(134), 77–97.

Ulwatunnisa, M., Retnawati, H., Muhardis, M., & Yusron, E. (2023). Revealing the characteristics of Indonesian language test used in the national-standardized school examinations. *REID (Research and Evaluation in Education)*, *9*(2), 210–222. https://doi.org/10.21831/reid.v9i2.31999

Waller, N. G., & Feuerstahler, L. (2017). Bayesian modal estimation of the four-parameter item response model in real, realistic, and idealized data sets. *Multivariate Behavioral Research*, *52*(3), 350–370. https://doi.org/10.1080/00273171.2017.1292893

Willse, J. T. (2018). *CTT: Classical test theory functions* (R package version 2.3.3) [Computer software]. https://cran.r-project.org/package=CTT

Yen, Y.-C., Ho, R.-G., Laio, W.-W., Chen, L.-J., & Kuo, C.-C. (2012). An empirical evaluation of the slip correction in the four parameter logistic models with computerized adaptive testing. *Applied Psychological Measurement*, *36*(2), 75–87. https://doi.org/10.1177/0146621611432862

Yim, L. W. K., Lye, C. Y., & Koh, P. W. (2024). A psychometric evaluation of an item bank for an English reading comprehension tool using Rasch analysis. *REID (Research and Evaluation in Education)*, *10*(1), 18–34. https://doi.org/10.21831/reid.v10i1.65284

Citation:
Santoso, A., Retnawati, H., Pardede, T., Apino, E., Rafi, I., Rosyada, M., Kassymova, G., & Wenxin, X. (2024). From investigating the alignment of *a priori* item characteristics based on the CTT and four-parameter logistic (4-PL) IRT models to further exploring the comparability of the two models. *Practical Assessment, Research, & Evaluation*, 29(14). Available online: https://doi.org/10.7275/pare.2043

Corresponding Author:
Agus Santoso
Universitas Terbuka
Jalan Cabe Raya, Pondok Cabe, Pamulang, Tangerang Selatan, Banten, Indonesia
Email: aguss@ecampus.ut.ac.id

Appendix 1.

The level of item difficulty justified based on the indicators and the items

Indicator	Item order	Difficulty level	Number of items
Explain economic terms starting with the letter A	1, 2, 3	Medium	3
Explain economic terms starting with the letter B	4, 5	Medium	2
Explain economic terms starting with the letter C	6, 7, 8	Easy	3
Explain economic terms starting with the letter D	9,10	Hard	2
Explain economic terms starting with the letter E	11, 12, 13	Medium	3
Explain economic terms starting with the letter F	14, 15	Hard	2
Explain economic terms starting with the letter G	16, 17, 18, 19	Medium	4
Explain economic terms starting with the letter I	20	Hard	1
Explain economic terms starting with the letter L	21, 22	Medium	2
Explain economic terms starting with the letter M	23, 24	Easy	2
Explain economic terms starting with the letter N	25	Medium	1
Explain economic terms starting with the letter O	26, 27	Medium	2
Explain economic terms starting with the letter P	28, 29	Medium	2
Explain economic terms starting with the letter Q or R	30	Hard	1

Appendix 2.

The estimation results of the parameter difficulty index and their labels based on CTT and IRT

Item ID.	*A priori*	CTT Difficulty index	Label	IRT Location (*b*)	Label
Item.1	Medium	0.784	Easy	–1.451	Easy
Item.2	Medium	0.774	Easy	–0.744	Medium
Item.3	Medium	0.721	Easy	–0.524	Medium
Item.5	Medium	0.428	Medium	1.012	Hard
Item.6	Easy	0.652	Medium	–0.780	Medium
Item.7	Easy	0.787	Easy	–0.956	Medium
Item.8	Easy	0.704	Easy	–0.659	Medium
Item.9	Hard	0.285	Hard	1.253	Hard
Item.10	Hard	0.642	Medium	0.011	Medium
Item.11	Medium	0.852	Easy	–1.316	Easy
Item.12	Medium	0.396	Medium	1.901	Hard
Item.13	Medium	0.929	Easy	–1.501	Easy
Item.14	Hard	0.263	Hard	1.041	Hard
Item.15	Hard	0.734	Easy	–0.637	Medium
Item.17	Medium	0.737	Easy	–1.032	Easy
Item.18	Medium	0.646	Medium	–0.607	Medium
Item.19	Medium	0.868	Easy	–1.998	Easy
Item.20	Hard	0.712	Easy	–0.582	Medium
Item.21	Medium	0.446	Medium	0.559	Medium
Item.22	Medium	0.517	Medium	0.343	Medium
Item.23	Easy	0.559	Medium	–0.575	Medium
Item.24	Easy	0.428	Medium	0.702	Medium
Item.25	Medium	0.188	Hard	2.611	Hard
Item.27	Medium	0.288	Hard	–0.369	Medium
Item.28	Medium	0.445	Medium	0.291	Medium
Item.29	Medium	0.700	Medium	–1.207	Easy
Item.30	Hard	0.812	Easy	–1.100	Easy

Appendix 3.

The point-biserial correlation coefficient (r_{pbis}) and slope parameter (a) for each item

Item ID.	CTT		IRT					
	r_{pbis}	Label of r_{pbis}	Slope (a)	Label of a	c	Label of c	u	Label of u
Item.1	0.328	Good	1.262	Fair	0.046	Acceptable	0.962	Acceptable
Item.2	0.491	Good	2.314	Good	0.183	Acceptable	1.000	Acceptable
Item.3	0.437	Good	1.828	Good	0.220	Acceptable	0.993	Acceptable
Item.5	0.248	Fair	3.267	Good	0.303	Unacceptable	0.971	Acceptable
Item.6	0.422	Good	1.580	Good	0.002	Acceptable	0.928	Acceptable
Item.7	0.431	Good	2.244	Good	0.173	Acceptable	0.966	Acceptable
Item.8	0.326	Good	2.005	Good	0.272	Unacceptable	0.897	Unacceptable
Item.9	0.232	Fair	5.757	Good	0.192	Acceptable	0.991	Acceptable
Item.10	0.406	Good	2.094	Good	0.291	Unacceptable	0.997	Acceptable
Item.11	0.418	Good	1.863	Good	0.124	Acceptable	0.999	Acceptable
Item.12	0.136	Fair	1.265	Fair	0.306	Unacceptable	0.999	Acceptable
Item.13	0.275	Fair	6.558	Good	0.389	Unacceptable	0.973	Acceptable
Item.14	0.280	Fair	4.182	Good	0.135	Acceptable	0.892	Unacceptable
Item.15	0.409	Good	6.071	Good	0.296	Unacceptable	0.897	Unacceptable
Item.17	0.319	Good	4.003	Good	0.222	Acceptable	0.845	Unacceptable
Item.18	0.457	Good	1.508	Good	0.002	Acceptable	0.983	Acceptable
Item.19	0.278	Fair	1.303	Fair	0.047	Acceptable	0.974	Acceptable
Item.20	0.444	Good	1.655	Good	0.163	Acceptable	0.999	Acceptable
Item.21	0.390	Good	1.735	Good	0.154	Acceptable	0.999	Acceptable
Item.22	0.442	Good	2.761	Good	0.213	Acceptable	0.999	Acceptable
Item.23	0.353	Good	1.573	Good	0.083	Acceptable	0.814	Unacceptable
Item.24	0.219	Fair	3.242	Good	0.305	Unacceptable	0.761	Unacceptable
Item.25	0.076	Poor	1.746	Good	0.159	Acceptable	0.993	Acceptable
Item.27	0.188	Fair	3.293	Good	0.127	Acceptable	0.384	Unacceptable
Item.28	0.302	Good	2.527	Good	0.222	Acceptable	0.771	Unacceptable
Item.29	0.338	Good	1.616	Good	0.028	Acceptable	0.872	Unacceptable
Item.30	0.319	Good	2.996	Good	0.329	Unacceptable	0.913	Acceptable