REID (Research and Evaluation in Education), 9(1), 2023, 86-117
Available online at: http://journal.uny.ac.id/index.php/reid

Gaining a deeper understanding of the meaning of the carelessness parameter in the 4PL IRT model and strategies for estimating it

Timbul Pardede[1*]; Agus Santoso[1]; Diki[1]; Heri Retnawati[2]; Ibnu Rafi[2]; Ezi Apino[2]; Munaya Nikma Rosyada[2]
[1]Universitas Terbuka, Indonesia
[2]Universitas Negeri Yogyakarta, Indonesia
[*]Corresponding Author. E-mail: timbul@ecampus.ut.ac.id

ARTICLE INFO

Article History
Submitted:
26 June 2023
Revised:
30 June 2023
Accepted:
30 June 2023

Keywords
carelessness parameter; dichotomous IRT; four-parameter logistic model; item response theory

Scan Me:

ABSTRACT

Three popular models are used to describe the characteristics of the test items and estimate the ability of examinees under the dichotomous IRT model, namely the one-, two-, and three-parameter logistic models. The three-item parameters are discriminating power, difficulty, and pseudo-guessing. In the development of the dichotomous IRT model, carelessness or upper asymptote parameter was proposed, which forms a four-parameter logistic (4PL) model to accommodate a condition where a high-ability examinee gives an incorrect response to a test item when he/she should be able to respond to the test item correctly. However, the carelessness parameter and the 4PL model have not been widely accepted and used due to several factors, and people's understanding of that parameter and strategies for estimating it is still inadequate. Therefore, this study aims to shed light on ideas underlying the 4PL model, the meaning of the carelessness parameter, and strategies used to estimate that parameter based on the extant literature. The focus of this study was then extended to demonstrating practical examples of estimating item and person parameters using the 4PL model using empirical data on responses of 1,000 students from the Indonesia Open University (Universitas Terbuka) on 21 of 30 multiple-choice items on the Business English test, a paper-and-pencil test. We mainly analyzed empirical data using the 'mirt' package in RStudio. We present the analysis results coherently so that IRT users would have a sufficient understanding of the 4PL model and the carelessness parameter, and they can estimate item and person parameters under the 4PL model.

To cite this article (in APA style):
Pardede, T., Santoso, A., Diki, D., Retnawati, H., Rafi, I., Apino, E., & Rosyada, M. (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. doi:https://doi.org/10.21831/reid.v9i1.63230

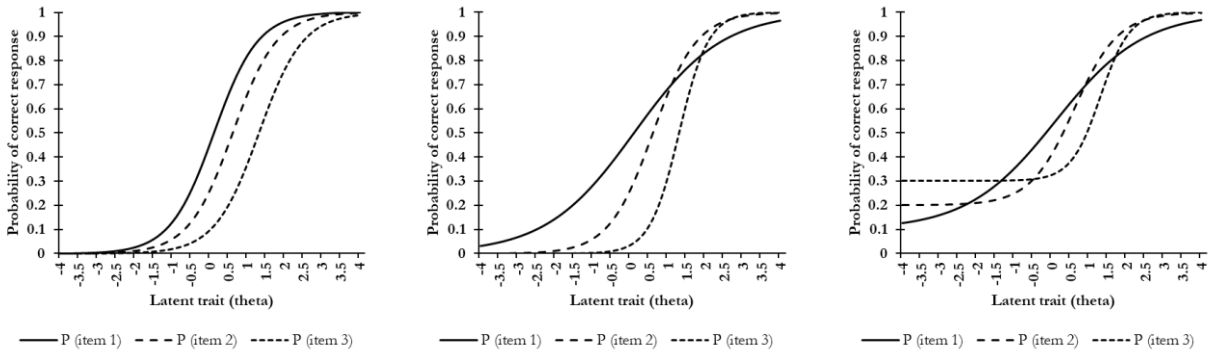
INTRODUCTION

It has been recognized that the classical test theory (CTT) approach, which is based on a linear model between the observed score and the sum of the true score and the error, has been used for a long time and is widely used to date to identify the characteristics of test items and the test itself because of the ease of calculation and interpretation it offers. However, the CTT approach has some drawbacks, one of which is that it is dependent on the sample or examinee group (Hambleton et al., 1991; Hambleton & Jones, 1993; Magno, 2009; Retnawati, 2016; Santoso et al., 2022; Zanon et al., 2016). This dependence suggests that a test item can be easy when the examinee has high ability. However, when the test items are presented to examinees who have low abilities, the test item may become difficult. Consequently, it would be difficult to obtain information on the characteristics of the test items accurately when we analyze the characteristics of the test using the CTT approach. Likewise for the case when we investigate students' abilities through analysis using the CTT approach, we cannot identify with certainty the extent of

ISSN: 2460-6995 (Online)

students' abilities because students' abilities on the test depend on the level of difficulty of the test items. Students obtain a high score on a test could be because the student's ability is high, or it could be because the items on the test are actually easy. Thus, with the CTT approach, it will be very difficult for us to identify the characteristics of a test and the test items with certainty because these characteristics will change with changes in the context of examinees, and to identify the characteristics of the examinees with certainty because these characteristics will change with changes in the context of the administered test (Hambleton et al., 1991).

In responding to a number of shortcomings of test item and test analysis with the CTT approach, modern test theory (also known as latent trait theory or item response theory (IRT)) was proposed. IRT, as an alternative theory test to CTT, provides some fundamental features which include the characteristics of the test items independent of the examinee group, the characteristics (i.e., abilities) of the examinees are independent of the test, the unit of analysis in the model is the test items rather than the test, reliability is not based on the parallel tests concept, and the ability of the examinees can be estimated precisely based on a non-linear model instead of a linear model as the CTT assumes (Bulut, 2015; Desjardins & Bulut, 2018; Hambleton et al., 1991; Hambleton & Swaminathan, 1985; Meijer & Tendeiro, 2018; Paek & Cole, 2020). The non-linear model on IRT represents a monotonically increasing function, called an item characteristic curve (ICC) or item characteristic function (ICC), for the probability of an examinee with a certain ability or proficiency (known as latent trait, θ) to provide the correct response to a test item that has certain characteristics (see Figure 1). In ICC or ICF, person and item characteristics are presented on the same scale, that is the logit scale. Certain characteristics attached to a test item, specifically for dichotomous or binary items, are referred to as item parameters, in which the number of item parameters represents the models in IRT. Item parameters under the IRT model which are usually the focus of studies consist of item discrimination (typically denoted by a), item location or difficulty (typically denoted by b), and guessing, pseudo-guessing, or pseudo-chance-level (typically denoted by c, or g in the 'mirt' package; hereafter, in this paper, we use the letter g to denote pseudo-guessing parameter).



(a) (b) (c)

Figure 1. Three ICCs from: (a) the 1PL IRT Model (Item 1: $b_1 = 0.1$; Item 2: $b_2 = 0.6$; and Item 3: $b_3 = 1.3$); (b) the 2PL IRT Model (Item 1: $a_1 = 0.5$, $b_1 = 0.1$; Item 2: $a_2 = 1$, $b_2 = 0.6$; and Item 3: $a_3 = 1.5$, $b_3 = 1.3$); and (c) the 3PL IRT Model (Item 1: $a_1 = 0.5$, $b_1 = 0.1$, $g_1 = 0.1$; Item 2: $a_2 = 1$, $b_2 = 0.6$, $g_2 = 0.2$; and Item 3: $a_3 = 1.5$, $b_3 = 1.3$, $g_3 = 0.3$)

In IRT, a model that represents the probability of an examinee j to correctly answer an item i that is dichotomously scored is determined by the ability of that examinee (θ_j) and the three item parameters is known as the three-parameter logistic (3PL) model, and the model is formulated by Birnbaum (1968) and we restate the model in Equation (1).

$$P_j\left(X_{ij}=1\middle|\theta_j;a_i,b_i,g_i\right)=g_i+\frac{1-g_i}{1+\exp\left[-Da_i\left(\theta_j-b_i\right)\right]}=g_i+\left(1-g_i\right)\frac{\exp\left[Da_i\left(\theta_j-b_i\right)\right]}{1+\exp\left[Da_i\left(\theta_j-b_i\right)\right]} \quad \ldots\ldots\ldots (1)$$

In Equation (1), D is a scaling constant which can be set to equal to 1 (when it is chosen to maintain the parameters at the scale of the logistic model) logistic function) or 1.702 (when it is chosen to represent the parameters into the same scale as the normal ogive model's parameters) (DeMars, 2010; Desjardins & Bulut, 2018; Hambleton et al., 1991; Hambleton & Swaminathan, 1985), b_i is the difficulty or location parameter of item i that represents the value of ability (θ) required by the examinee so that the probability of answering the item i correctly is 0.5, a_i is the discrimination or slope parameter of item i that represents the degree to which item i differentiates examinees based on their latent traits or abilities which is consistent with the steepness of the slope of the tangent to the ICC of item i at $\theta = b_i$, and g_i is pseudo-guessing parameter of item i that represents the probability of examinees whose abilities are at a low level to be able to answer item i correctly or lower asymptote of the ICC of item i. When the 3PL model is reduced by adding certain constraints to one or two item parameters, we can obtain 1PL and 2PL models. The 2PL model is a special case of the 3PL model which is formed when we set $g_i = 0$. When we set $a_i = a$ (item discrimination for all items has the same value, i.e., a constant a) and $g_i = 0$ in the 3PL model, we obtain the 1PL model. Furthermore, when the constant a in the 1PL model is set equal to 1, we obtain the Rasch model. Although both the Rasch and 1PL models involve only one item parameter, namely b, which leads them to be considered the same on many occasions for practical purposes, it is better not to consider the two models the same because they have different underlying concepts (see Andrich (2004) for further review on this topic). To provide an overview of the relationship between item and person parameters in the three models in the dichotomous IRT, we provide ICCs of the three items with various parameter estimates according to the model used to estimate the parameters as presented in Figure 1.

Similar to the CTT which was developed under a number of assumptions (see Allen & Yen, 1979), IRT was also developed with a number of assumptions that must be satisfied to ensure the advantages of using IRT over CTT, one of which is in terms of accurately estimating item and person parameters. There are three major assumptions underlying IRT, namely unidimensionality, parameter invariance, and local independence (DeMars, 2010; Hambleton et al., 1991; Hambleton & Swaminathan, 1985; Retnawati, 2014, 2016). The strict unidimensionality assumption requires that there is only a single latent trait (θ) that is measured in a test for each examinee that influences their performance in the test or responding test items. Because it cannot be denied that there are other factors that also influence the examinee's performance on a test even though they can be ignored and can be considered as random errors (DeMars, 2010), instead of only emphasizing the existence of a single latent trait or ability, unidimensionality in IRT is then more directed at the essentially unidimensionality (Slocum-Gori & Zumbo, 2011) which is sufficient to require presence of a 'dominant' component or factor, namely θ, which has an impact on item responses or test performance of examinees (Hambleton et al., 1991; Hambleton & Swaminathan, 1985; Retnawati, 2014, 2016; Slocum-Gori & Zumbo, 2011). The next assumption is parameter invariance which includes item parameter invariance and person parameter invariance. Item parameter invariance requires that when a group of examinees is divided into two subgroups based on certain conditions, the values of the item parameter estimates of the two subgroups must be similar or tend to be similar. Meanwhile, person parameter invariance suggests that when the test items are divided into two subgroups based on certain conditions, the estimation of the person parameter (ability) using the item responses from the two subgroups must be similar or tend to be similar. The last major assumption of IRT is local independence, which this assumption requires that the examinee's performance in responding to one test item is not related to or not affected by the examinee's performance in responding to other test items.

Given the underlying assumptions of IRT and the advantages that IRT has over CTT, the IRT approach has been used in item response analysis in various large-scale assessments, such as national examinations (Bulut, 2015; Doğruöz & Arikan, 2020), and suggests that 2PL and 3PL models are more favored and frequently used to estimate item parameters and ability of examinee in the large-scale assessment (Bulut, 2015; Desjardins & Bulut, 2018; Doğruöz & Arikan, 2020; Rutkowski et al., 2014). It has been considered as common to use multiple-choice items in large-scale assessments because of several advantages that this type of test item offers which include ease of scoring so as to save time and costs, the scores obtained from these item types are more reliable, and a broader range of topics can be covered (DiBattista & Kurzawa, 2011; Haladyna & Rodriguez, 2013; Quaigrain & Arhin, 2017; Rafi et al., 2023). However, the use of multiple-choice items in some conditions faces challenges, one of them is the opening of a large opportunity for examinees to guess randomly or guess with certain considerations the extant choices in responding to the test item, so that the examinee's response to the test item does not reflect their true competence on that test item accurately. The challenge that arises from using multiple-choice items in a test due to the possibility of guessing leading to the correct response (lucky guessing) should be taken into account in statistical modeling (Haladyna et al., 2019; Kubinger et al., 2010). The 3PL IRT model, therefore, has addressed that challenge by involving a pseudo-guessing parameter represented by the lower asymptote of the ICC such that it is possible for us to model the relationship between the item parameters and the examinee's ability.

When the items used in a test are of the multiple-choice type, there is a possibility that a high ability examinee who should be able to give the correct response to a test item may actually respond incorrectly to the test item. Responses obtained from such conditions are considered as aberrant responses and are considered to be a threat to accurately identify the examinee's actual competence or ability (Liao et al., 2012). Even further Liao et al. (2012) argued that an aberrant response from a high ability examinee is more dangerous, especially if it occurs at the beginning of the test or in the first few items of the test, compared to an aberrant response from random or strategic guessing behavior that results in the correct answer by a low-ability examinee. The 3PL model certainly cannot account for the aberrant responses that high ability examinees exhibit and their ability estimates tend to be underestimated under the model (Liao et al., 2012), so Barton and Lord (1981) have long proposed an idea that could accommodate the presence of such aberrant responses in order to produce a more accurate estimate of examinee's ability. The idea that Barton and Lord (1981) put forward, which became known as the four-parameter logistic (4PL) model, leads to the possibility that the value of the upper asymptote in the ICC of an item is less than 1, but remains as close as possible to 1.

Although the 4PL model was introduced more than three decades ago, it encounters a number of theoretical and practical hurdles which make it less attention, less explored, and less widely used than the other three IRT models and Rasch model. One of the theoretical and practical obstacles to the 4PL model is shown by the Barton and Lord's (1981) study which shows that setting the upper asymptote parameter values to be not equal to 1, which in this study set the upper asymptote values to be 0.99 and 0.98, it turns out that in general did not show a change in the examinee's ability estimate compared to when the upper asymptote parameter was made equal to 1 (i.e., 3PL model); later Primi et al.'s (2018) study also found the similar results. Because of the study results they obtained, they suggested that the 4PL model should not be urgent to use, moreover due to the limitations of available technology, it takes a lot of time to perform complex computations and derivations, and it is also difficult to estimate the parameter, as pointed out by several studies (e.g., Kalkan, 2022; Kalkan & Çuhadar, 2020; Liao et al., 2012; Loken & Rulison, 2010) based on the extant literature, when working with this model. Based on the findings of Barton and Lord's (1981) study, Hambleton and Swaminathan (1985) argued that the 4PL model is only attractive for theoretical reasons, not for practical.

Along with the development of technology that allows for complex computations, a number of studies (e.g., Doğruöz & Arikan, 2020; Kalkan, 2022; Liao et al., 2012; Loken & Rulison,

2010; Ogasawara, 2017; Primi et al., 2018; Robitzsch, 2022; Waller & Feuerstahler, 2017; Yen et al., 2012) have been carried out to further explore the 4PL model at both theoretical and practical levels based on empirical or simulations studies. Although it cannot be denied that the 4PL model still has a number of challenges so that it still needs improvement, those studies indicate the promising potential of the 4PL model in practice, especially in the computerized adaptive testing (CAT), in terms of estimating the ability of examinees efficiently and with precision when there is a high ability examinee exhibiting an aberrant response due to careless behavior (see Cheng & Liu, 2015; Liao et al., 2012; Rulison & Loken, 2009; Yen et al., 2012). Given that high ability examinees' careless behavior in responding to a test item also likely to occur in a paper-and-pencil testing, our study was intended to demonstrate the use of the 4PL IRT model to estimate item and ability parameters under a paper-and-pencil testing and empirical data drawn from the testing by using 'mirt' package (Chalmers, 2012) under RStudio (Posit Team, 2023) working environment which can be accessed and used for free. The purpose of our study is in line with what previous studies (e.g., Kalkan, 2022; Liao et al., 2012; Magis, 2013; Rulison & Loken, 2009) recommended that further studies need to promote practical examples of using the 4PL model with empirical data because it can become an additional basis for further use and development of the 4PL model. In addition, based on what Loken and Rulison (2010) have suggested that studies that focus on parameter estimation and their interpretation in the 4PL model still need to be undertaken more, the purpose of this study was thus extended to shed light on the meaning of the upper asymptote or carelessness parameter and the 4PL model itself.

METHOD

Study Design and Data Collection

This descriptive study focuses on (1) describing the meaning of carelessness or slipping parameter in the 4PL IRT model and strategies for estimating this parameter through literature reviews and (2) describing the characteristics of a test and its items based on based on the 4PL IRT model through the use of empirical data along with demonstrating how to estimate item and person parameters in that IRT model. The second focus of this study was carried out using empirical data on the responses of 5000 students from the *Universitas Terbuka* (UT; Indonesia Open University) in the *Bahasa Inggris Niaga* (or Business English) test. This test is a paper-and-pencil test with 30 four-option multiple-choice questions. Student responses to the test were scored dichotomously (i.e., 1 = correct answer and 0 = incorrect answer or no response). Since the second focus of this study was emphasized more on demonstrating the estimation of item and person parameters based on the 4PL IRT model, the data used in this study consisted of dichotomous scores from the responses of 1000 students who were randomly selected from 5000 students to the first 21 test items out of 30 existing test items (M = 12.597, SD = 3.793). Later, the data that we used in the analysis with the IRT approach were the scores obtained by 1000 students on 19 test items (M = 12.044, SD = 3.829).

Data Analysis

In this section we describe how we conducted a quantitative analysis of empirical data on student response scores on the Business English test to reveal item and person parameters based on the 4PL IRT model. First, we analyzed the test items using the CTT approach with a focus on detecting test items that failed to distinguish students with high abilities from those with low abilities in the test as indicated by the point-biserial correlation coefficient (r_{pb}) of these items which are non-positive. We performed this analysis in RStudio (Posit Team, 2023), an integrated development environment (IDE) for R, using the 'CTT' package (Willse, 2018). The R script we used to investigate the characteristics of the test and its items based on CTT approach (i.e., test score reliability based on Cronbach's alpha, standard error of measurement (SEM), and the dis-

criminating power of item based on the point-biserial correlation coefficient) is available in the Appendix 1. We found that of the 21 existing test items ($\alpha = 0.761$, $SEM = 1.853$), two items had a negative value of r_{pb}, namely item 10 ($r_{pb} = -0.076$) and item 17 ($r_{pb} = -0.158$). Thus, excluding these two items in the subsequent analysis, there are 19 test items ($\alpha = 0.793$, $SEM = 1.743$) that we used in conducting the analysis using the IRT approach. The entire analysis using the IRT approach was carried out in RStudio (Posit Team, 2023).

In the analysis using the IRT approach, the first step we performed was to test the fit of student responses to the test items with the Rasch, 1PL, 2PL, 3PL, and 4PL models. We carried out this test, called the item fit test or the goodness-of-fit test of items, using the 'mirt' package (Chalmers, 2012), where with this package, the item fit test is assessed by default using the signed chi-squared (s-χ^2) statistics proposed by Orlando and Thissen (2000, 2003), where an item is said to be fit with the model under investigation when the p-value that the statistic is not statistically significant ($p \geq 0.05$). The more items that fit a model, the model is more favored than the others for estimating item and person parameters. Afterwards, we carried out overall model-fit test (also known as goodness-of-fit test of the model) to examine which model that fit the data. The overall model-fit test was performed with a number of model-fit indices, such as M_2, Akaike information criterion (AIC), Bayesian information criterion (BIC), root mean square error of approximation (RMSEA), standardized root mean square residual (SRMSR), log-likelihood, and χ^2. We used the results from the item-fit and model-fit tests in the person-fit test to examine the appropriateness of the pattern of responses each examinee gives to the existing test items to the model of interest using z_h statistics (Drasgow et al., 1985). After we performed item-fit, model-fit, and person-fit tests, we estimated the item parameters with the 'mirt' package. By using this package, item parameters estimation was carried out using the fixed quadrature expectation-maximization (EM) algorithm (see Chalmers, 2012) as the default method, although there we may use quasi-Monte Carlo EM, Monte Carlo EM, Stochastic EM, Metropolis-Hastings Robbins-Monro (MH-RM), and Bock and Lieberman approach (Chalmers, 2023). Along with estimating item parameters, we also provided item characteristic curve (ICC) and item information function (IIF) for each test item under investigation in this study. Under the model of interest, we estimated the ability or person parameter based on three methods as available in the 'mirt' package, i.e., maximum likelihood (ML), maximum a posteriori (MAP), and expected a posteriori (EAP). We provided results regarding test information function (TIF) and standard error of measurement (SEM) along with empirical and marginal reliability estimates. The R syntax that we used to estimate item and person parameters as well as identify test characteristics based on TIF and SEM under IRT model is available in Appendix 3.

Lastly, to ensure the accuracy of item and person parameters estimates under IRT, we provided evidence of the fulfillment of the three major assumptions of IRT model, namely unidimensionality, parameter invariance, and local independence. We examined the assumption of unidimensionality through factor analysis and principal components using the 'psych' package (Revelle, 2023) and by considering the various indices and rules in determining the number of dominant factors that have been summarized by Hattie (1985) and Slocum-Gori and Zumbo (2011). We then used Pearson's correlation coefficient to demonstrate that the item and person parameter estimates of the two subgroups are identical or not much different. By using the 'mirt' package (Chalmers, 2012), it is possible for us to detect local dependence of a pair of items with four statistics, namely Yen's Q_3 (Yen, 1984), χ^2 (Chen & Thissen, 1997), likelihood-ratio (G^2) (Chen & Thissen, 1997), and jack-knife slope index (JSI) (Edwards et al., 2018). The R syntax that we used to test the adequacy of the major assumptions of the IRT model is provided in Appendix 2.

FINDINGS AND DISCUSSION

In this section, we report the two main findings of this study, namely the meaning of the carelessness or slipping parameter and the results of item and person parameters estimation based on the 4PL IRT model. The demonstration of the meaning of the carelessness parameter is

integrated with the demonstration of the basic concepts of the 4PL IRT model and various strategies or methods that previous studies have suggested or used to estimate the carelessness parameter. An explanation of the results of item and person parameters estimation based on the 4PL IRT model is presented systematically so that it can demonstrate procedures for estimating these parameters, including examining the adequacy of the assumptions within the IRT framework and estimating measurement precision.

The Four-Parameter Logistic (4PL) IRT Model and the Carelessness Parameter

While the 3PL model was developed from the 2PL model to accommodate the aberrant response of low-ability examinees, as we mentioned earlier, the 4PL model was proposed to address the aberrant response of those with high ability. The proposed 4PL model basically considers that the upper asymptote of an ICC is not equal to 1, which is different from what we find in an ICC generated under the 3PL model. Consideration of the upper asymptote estimate value which is lower than 1 is to compensate for an incorrect response that the examinee should not give to an item and to remain in line with the understanding that high ability examinees still have a greater probability of responding to an item test correctly than those with low ability. The extant literature suggests that when we estimate item and person parameters under the 3PL model, a high ability examinee who responds incorrectly to an item due to some factors will have a zero (0) probability of responding to the test item correctly (Doğruöz & Arikan, 2020; Loken & Rulison, 2010; Magis, 2013); and thus, the 4PL model does not allow this sort of thing to happen. Furthermore, this sort of thing would lead to an underestimation of the true ability of the high ability examinee, moreover the aberrant responses exhibited by the high ability examinees are certainly very rare compared to the aberrant responses that the low-ability examinees made through random or strategic guessing. We have mentioned earlier that there are several possible factors that could cause a high ability examinee to answer the wrong test item when he/she should be able to answer it correctly. These possible factors include fatigue, anxiety or stress, unfamiliarity with the mode or condition of how a test is administered, carelessness in reading questions, inattentiveness in responding to an item, and being unmotivated to work on questions that are too easy or not challenging (Antoniou et al., 2022; Liao et al., 2012; Magis, 2013; Rulison & Loken, 2009). Accordingly, the terms inattention parameter (Doğruöz & Arikan, 2020; Magis, 2013), carelessness parameter (Antoniou et al., 2022; Barnard-Brak et al., 2018), and slipping parameter (Kalkan & Çuhadar, 2020; Robitzsch, 2022) are commonly used in some literature to refer to the upper asymptote parameter.

$$
\begin{aligned}
P_j\left(Y_{ij}=1\middle|\theta_j;a_i,b_i,g_i,u_i\right) &= P_j\left(Y_{ij}=1\cap X_{ij}=0\right)+P_j\left(Y_{ij}=1\cap X_{ij}=1\right)\\
&= P_j\left(Y_{ij}=1\middle|X_{ij}=0\right)P_j\left(X_{ij}=0\right)+P_j\left(Y_{ij}=1\middle|X_{ij}=1\right)P_j\left(X_{ij}=1\right)\\
&= g_i\left[1-P_j\left(X_{ij}=1\right)\right]+u_iP_j\left(X_{ij}=1\right) \qquad \text{..... (2)}\\
&= g_i\left[1-\frac{\exp\left[Da_i\left(\theta_j-b_i\right)\right]}{1+\exp\left[Da_i\left(\theta_j-b_i\right)\right]}\right]+u_i\left[\frac{\exp\left[Da_i\left(\theta_j-b_i\right)\right]}{1+\exp\left[Da_i\left(\theta_j-b_i\right)\right]}\right]\\
&= g_i+\left(u_i-g_i\right)\frac{\exp\left[Da_i\left(\theta_j-b_i\right)\right]}{1+\exp\left[Da_i\left(\theta_j-b_i\right)\right]}
\end{aligned}
$$

A number of literature (e.g., Antoniou et al., 2022; Doğruöz & Arikan, 2020; Kalkan, 2022; Kalkan & Çuhadar, 2020; Loken & Rulison, 2010; Magis, 2013; Waller & Feuerstahler, 2017) have provided a general form of a function (see Equation 2) which states the 4PL model based on what Barton and Lord (1981) have introduced. In this paper, we prefer to present the 4PL

model based on the ideas put forward by Battauz (2020) which derives the 3PL and 4PL models by not only considering the response that an examinee gives to a test item but also considering whether the examinee really knows the correct answer. Suppose that the first consideration is represented by the binary random variable Y and the latter is represented by the binary random variable X. When an examinee j gives the correct response to item i, $Y_{ij} = 1$, and when the response given is wrong, $Y_{ij} = 0$. Furthermore, when examinee j knows the correct response for item i, $X_{ij} = 1$, whereas when the examinee does not know the correct response for item i, $X_{ij} = 0$. Through these two random variables and the 2PL model based on one binary random variable X, the 3PL and 4PL models can be formulated with the understanding that examinee j with certain ability (θ_j) can give the correct response to item i because the examinee does know the correct response or probably the examinee does not actually know the correct response to that item. Accordingly, the probability that an examinee with ability θ gives the correct response to a test item is the sum of the probability that the examinee is able to respond to the item correctly but does not know that the response is the correct one (i.e., $P_j(Y_{ij} = 1 \cap X_{ij} = 0) = P_j(Y_{ij} = 1 \mid X_{ij} = 0)P_j(X_{ij} = 0)$) and the probability that the examinee is able to give the correct response and knows that it is the correct response for the item (i.e., $P_j(Y_{ij} = 1 \cap X_{ij} = 1) = P_j(Y_{ij} = 1 \mid X_{ij} = 1)P_j(X_{ij} = 1)$). By using the axioms of probability to demonstrate conditional probability to make it intuitively easier to understand and interpret, we can formulate the 4PL model as shown in Equation 2.

In Equation 2, the expression $P_j(Y_{ij} = 1 \mid X_{ij} = 0)$ represents the estimation of the pseudo-guessing parameter (lower asymptote, g_i) for item i. Meanwhile, the expression $P_j(Y_{ij} = 1 \mid X_{ij} = 1)$ in Equation 2 reflects the estimation of the carelessness parameter (upper asymptote; u_i) for item i. Although most of the literature prefers to use d or δ to denote the carelessness parameter, in this paper we prefer to use the letter u (upper asymptote) to denote the parameter so that it is consistent with what is used in and the output of the 'mirt' package. Thus, we can interpret the carelessness parameter as an estimate of the probability of an examinee with a high certain ability to be able to answer a test item correctly and know the correct answer for that item. The greater the estimated carelessness parameter, the greater the probability that a high ability examinee will provide the correct answer for an item because the examine does have sufficient knowledge to know the correct answer for an item and the higher the upper asymptote of the corresponding ICC. The smaller the parameter estimate, the smaller the probability for an examinee with high ability to correctly answer an item that he/she should be able to answer the item correctly based on that ability and the upper asymptote of the corresponding ICC. This latter condition indicates the increasing negligence or carelessness of the examinee in answering an item that he/she is actually able to answer correctly.

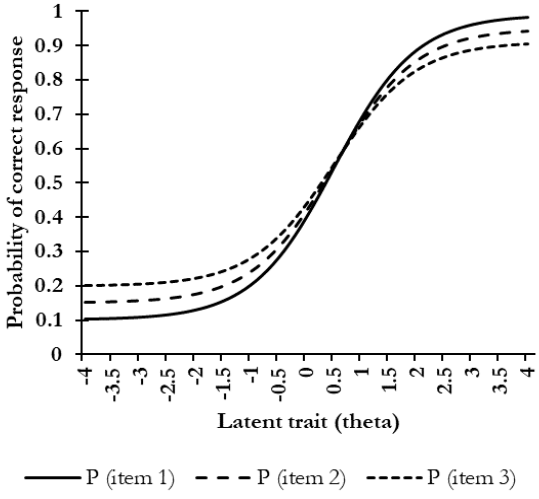


Figure 2. Three ICCs from the 4PL IRT Model ((Item 1: $a_1 = 0.8$, $b_1 = 0.5$, $g_1 = 0.1$, $u_1 = 0.99$; Item 2: $a_2 = 0.8$, $b_2 = 0.5$, $g_2 = 0.15$, $u_2 = 0.95$; and Item 3: $a_3 = 0.8$, $b_3 = 0.5$, $g_3 = 0.2$, $u_3 = 0.91$)

Figure 2 illustrates the three ICCs which represent the relationship between fluctuations in the estimated carelessness parameter and the probability of examinees with high abilities giving the correct responses to these items. The difference in the probability of high ability examinees to correctly answer an item will converge with the difference in the estimated carelessness parameter of each of these items. Although until now we have not obtained enough literature to set criteria for the extent to which the quality of an item is viewed from the carelessness parameter of the item, based on Figure 2 we argue that the closer to 1 the estimate of the carelessness parameter is better because it more closely reflects high ability examinees as they should. Barnard-Brak et al. (2018) suggest that the estimation of the carelessness parameter of an item should remain high, that is, it should be more than or equal to 0.9. Accordingly, we can say that in Figure 2, good items in terms of the estimation of the carelessness parameter, namely item 1, item 2, and item 3, respectively.

Several previous studies have attempted to identify and elaborate methods or algorithms that can overcome the challenges of using the maximum-likelihood method (e.g., difficult to estimate the parameters under the 4PL model and risk getting unreliable carelessness parameter estimates) (Loken & Rulison, 2010) and can be used to estimate the magnitude of the carelessness parameter. Kalkan (2022) in his study has identified a number of algorithms that can be used for parameters estimation consisting of expectation-maximization (EM), Monte Carlo EM (MCEM), quasi-Monte Carlo EM (QMCEM), Metropolis-Hastings Robbins-Monro (MH-RM), Bayesian modal estimation (BME), Gibbs sampler, and Markov chain Monte Carlo (MCMC). The last three algorithms work under the Bayesian framework. Previous studies have attempted to investigate which algorithm is best for estimating parameters under the 4PL model, and found that there is no best algorithm but rather that there is an algorithm that is better than the other algorithms under certain conditions (e.g., test length, sample size, number of factors) (see Kalkan, 2022; Loken & Rulison, 2010). By comparing the three algorithms (i.e., EM, QMCEM, and HR-RM) through a simulation study, Kalkan (2022) suggest that to estimate item parameters under 4PL model it is better to use EM algorithm when there is only one factor, while the MH-RM algorithm is better to use when there are two or three factors. Nonetheless, the three algorithms yielded similar estimates of carelessness parameters for the three conditions manipulated in his study (i.e., number of factors, the magnitude of the correlation between the factors, and test length or the number of test items). Furthermore, Loken and Rulison (2010) argued that the use of an algorithm based on the Bayesian framework might produce a good and reliable estimate of the carelessness parameter when the parameter is item-specific. However, the Bayesian framework can bring its own challenges in terms of sample size as shown by Waller and Feuerstahler (2017) that the BME algorithm yields accurate estimation of item parameters under the 4PL model when the sample size used is at least 5,000.

IRT Model Assumptions

The first assumption that has to be satisfied in IRT is the assumption of unidimensionality which requires that the Business English test only measures one dominant factor or latent trait (ability). This can be demonstrated through factor analysis and principal components. We ensured in advance that the data of student responses to the extant items were suitable for factor analysis and principal components by considering the correlation matrix and sample adequacy. The Bartlett's test of sphericity ($\chi^2(171) = 2570$, $p < 0.001$) suggested that there is sufficient evidence not to accept that the correlation matrix formed is an identity matrix. The Kaiser-Meyer-Olkin (KMO) factor adequacy (Overall MSA = 0.9; MSA for each item ranges from 0.63 to 0.93) demonstrated that that the sample of responses on the test items are sufficient for each variable in the model and the complete model. Based on these two results, we can proceed with the analysis on the extant item responses data to the factor analysis and principal components. The ratio of the first eigenvalue to the second eigenvalue (i.e., $\lambda_1/\lambda_2 = 3.47/0.52 = 6.67$ for factor analysis and $\lambda_1/\lambda_2 = 4.25/1.35 = 3.15$ for principal components), the number of factors or components

whose eigenvalues are greater than 1 based on factor analysis (see Figure 3), the explained variance of 22% for one component based on the principal components, and the eigenvalue for one factor or component that are in stark contrast to the eigenvalues for two or more factors or components (see Figure 3) demonstrate the existence of one dominant factor or latent trait. In other words, student responses to the items contained in the Business English test support the assumption of unidimensionality which requires that the differences in responses that students give to the items in the Business English test are mainly caused by one variable, ability, or latent trait related to Business English itself. Although it is possible that student responses to the test items are also influenced by other abilities or latent traits, their influence on student responses to the test items can thus be neglected.

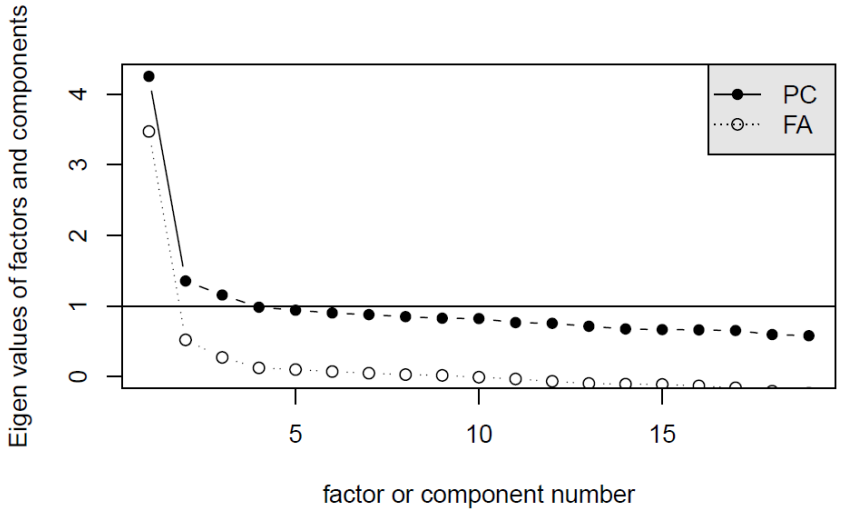


Figure 3. Scree Plot Representing the Eigenvalues by the Principal Components and Factor Analysis

The second assumption underlying IRT is parameter invariance. There are two levels of parameter that we need to prove that they are invariance, namely the item parameter and the person parameter. The item parameter invariance suggests that the test item parameters, in which the number of item parameters depends on which IRT model fits, is not affected by the diversity or distribution of examinee's abilities (Hambleton et al., 1991; Hambleton & Swaminathan, 1985). This implies that the parameter estimation results of an item will not change, tend to be identical, or not much different when the item is administered to different subgroups in a population or different contexts (Hambleton et al., 1991; Retnawati, 2014, 2016; Rupp & Zumbo, 2004). Because later it will be shown that the parameters of the items used in the Business English test are better fit to be estimated under the 4PL model, evidence of fulfilling the assumption of invariance of item parameters was provided based on four parameters, namely *a*, *b*, *g*, and *u*. We obtained evidence of the invariance of the four item parameters by forming two subgroups that were formed randomly from the examinee's data in this study with a proportion of 0.5 for each subgroup so that two subgroups with the same or nearly the same size were obtained. We then performed correlation analysis on the item parameter estimation results from each subgroup to investigate the strength of the relationship between the two parameter estimation results. The Pearson's correlation analysis demonstrates a strong positive relationship (Schober et al., 2018) between parameter *a* estimates ($r(17) = 0.715$, $p = 0.00059$), parameter *b* estimates ($r(17) = 0.904$, $p < 0.001$), parameter *g* estimates ($r(17) = 0.792$, $p < 0.001$), and parameter *u* estimates ($r(17) = 0.813$, $p < 0.001$) estimated from the first subgroup (Subgroup 1) and those estimated from the second subgroup (Subgroup 2). The scattergram showing the distribution of the estimated results for each item parameter estimated from the two subgroups is presented in Figure 4. Thus, the assumption of invariance on the item parameters has been satisfied.

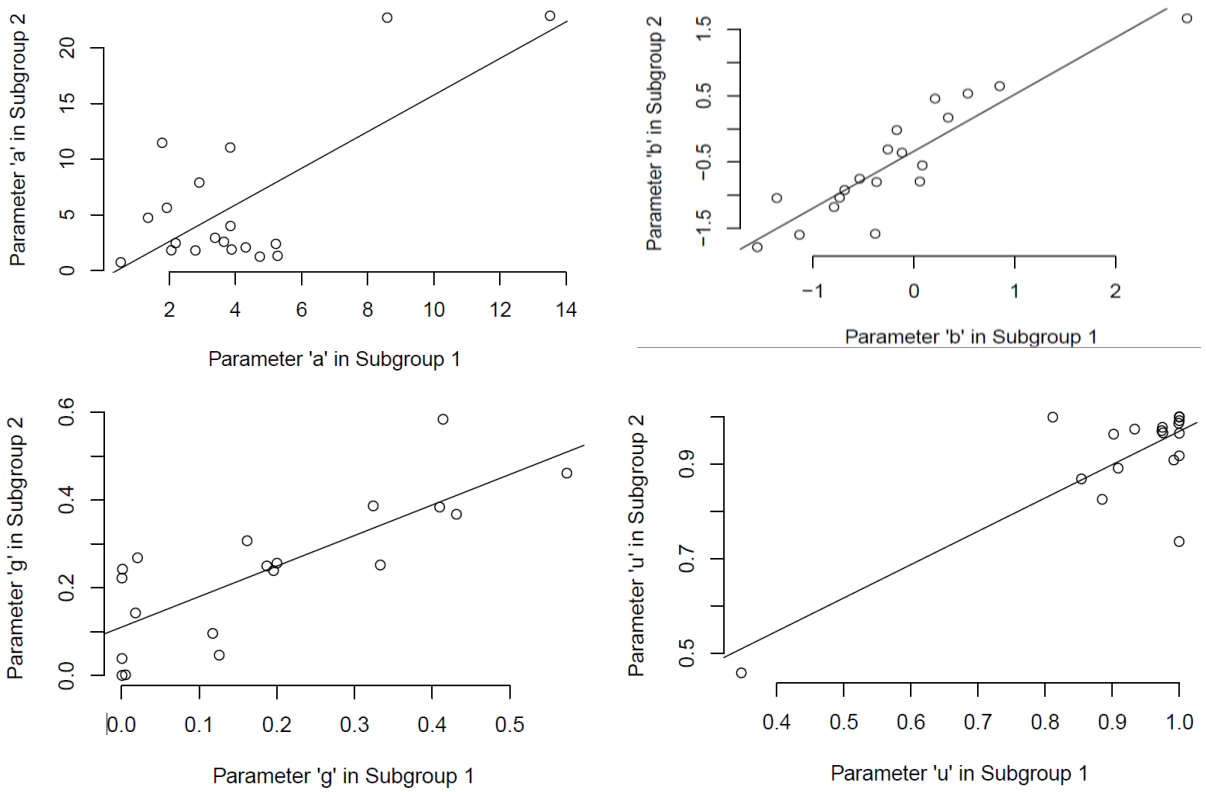


Figure 4. The Scattergram Showing the Distribution of Item Discrimination (*a*), Item Difficulty (*b*), Pseudo-guessing (*g*), and Carelessness (*u*) Estimates Estimated from the Two Different Subgroup

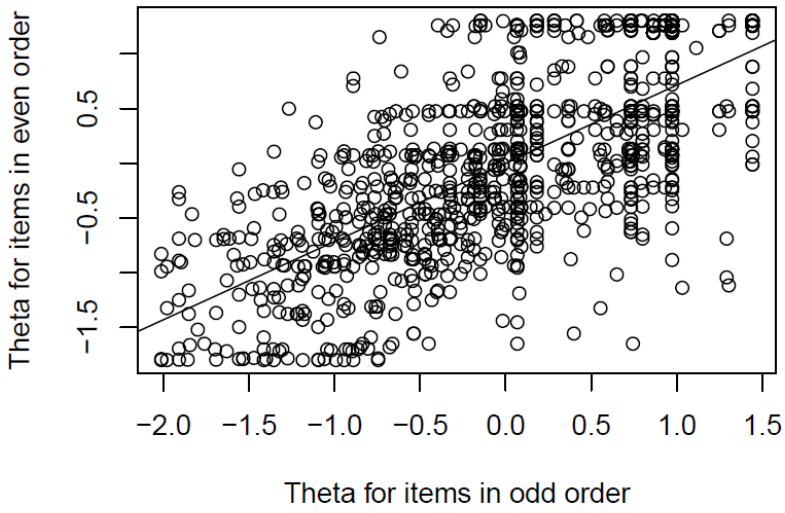


Figure 5. The Scattergram Showing the Distribution of Examinees' Abilities Estimated from a Subset of Items in Odd Order and a Subset of Items in Even Order

Using the same understanding as item parameter invariance, this ability or person parameter invariance refers to the consistency of the examinee's ability estimates obtained from two subsets of the items contained in the Business English test, in which two subsets of the test items can be formed based on certain conditions (Hambleton et al., 1991; Hambleton & Swaminathan, 1985; Rupp & Zumbo, 2004). The condition that we used in this study to form two subsets of test items is the order of the test items, where the first subset consists of odd order test items and the second subset consists of even order test items. By assuming that the items follow the 4PL model and by using expected a posteriori (EAP), the default factor score estimation method in

the 'mirt' package, to estimate the abilities of examinees, we obtained the distribution of examinees' abilities estimated based on items in odd order and based on those in even order (see Figure 5). Pearson's correlation analysis shows that there is a strong positive relationship (Schober et al., 2018) between examinee's ability estimates estimated from subset of items in odd order and those estimated from subset of items in even order ($r(998) = 0.701$, $p < 0.001$). Hence, the assumption of invariance on the person or ability parameter (θ) has also been satisfied.

The last IRT assumption that we need to show that it is satisfied is local independence, where this assumption requires that the examinee's response to one item is independent of his response to other items. In other words, this assumption implies that the good or bad response that the examinee gives to an item has no effect on the better or worse response he gives to other items. We provided evidence that the local independence assumption is satisfied through four statistics, namely Yen's Q_3, Person's χ^2, likelihood-ratio G^2, and JSI. Yen's Q_3 statistic in this study reflect the magnitude of the correlation between the residuals of each pair of items (there is 171 pairs of items in this study), where the residuals show the difference between the observed score (i.e., 0 or 1) and the predicted score based on the better fit dichotomous IRT model (i.e., $P(\theta)$). It has been suggested that the residuals of each pair of items are uncorrelated or that the correlation coefficient should be as close to zero as possible (Christensen et al., 2017). In addition, the absolute value of the Yen's Q_3 of a pair of items that is greater than 0.2 indicates violation of local independence (Chen & Thissen, 1997).

Two other statistics used to detect local dependence are Pearson's χ^2 and likelihood-ratio G^2, both of which are based on the relationship between the observed score frequencies and the predicted score frequencies through a particular dichotomous IRT model that best fits, expressed in the Pearson chi-square formula and the chi-square likelihood-ratio formula, respectively (Chen & Thissen, 1997). Through the 'mirt' package, these two statistics are presented in the form of signed Cramers V coefficients which are the standardized values of the residuals of a pair of items (Chalmers, 2023). Although there is no definite cut-off value to determine the existence of local dependence, we expect that the coefficient value for each pair of items is close to zero, which means that the existing pairs of items are close to being independent of each other. In this study, we set 0.3 as the cut-off (Chen & Thissen, 1997), where the absolute value of the standardized values of the residuals of a pair of items greater than 0.3 demonstrates local dependence. The last statistic that we also consider in diagnosing the presence of local dependence is the jack-knife slope index (JSI). JSI represents the change in the slope (discrimination) parameter of an item estimated from the full data with that estimated from the data without a response in another item relative to the standard error of the slope (discrimination) parameter from the estimate based on the full data (Edwards et al., 2018; Houts & Edwards, 2013). The absolute value of summed JSI of a pair of items that is greater than the sum of the mean and twice the standard deviation of summed JSI of all item pairs indicates that the item pair that the item pair is locally dependent (Edwards et al., 2018; Houts & Edwards, 2013).

Table 1. Summary Statistics of Local Dependence Detection

Statistics	Min.	Q1	Mdn	*M* (*SD*)	Q3	Max.
Yen's Q_3	–0.169	–0.075	–0.044	–0.042 (0.050)	–0.013	0.098
Pearson's χ^2	–0.122	–0.030	–0.009	–0.007 (0.040)	0.019	0.103
Likelihood-ratio G^2	–0.115	–0.030	–0.009	–0.006 (0.038)	0.019	0.105
Summed JSI	–0.900	–0.204	0.006	0.020 (0.359)	0.230	1.208

Note: Min. = Minimum value, Q1 = First quartile, Mdn = Median, Q3 = Third quartile, Max. = Maximum value, and JSI = Jack-knife slope index

Based on Table 1, there are no pairs of items that are detected to be locally dependent in terms of the Yen's Q_3 statistic because there are no pairs of items whose absolute value of the Yen's Q_3 is more than 0.2. This result is supported by the results of local dependency detection through Pearson's χ^2 and likelihood-ratio G^2 statistics which show that the absolute value of the

signed Cramers V coefficients of each item pair is not greater than 0.3 (see Table 1), meaning that all item pairs are locally independent. However, considering the results of the JSI statistic, we found that nine out of 171 pairs of items have an absolute value of the summed JSI of more than 0.738, meaning that the nine item pairs are locally dependent. Taking into account the overall results of the four statistics for detecting the presence of local dependence on a pair of items and the understanding that local dependence statistics are more emphasized for diagnostic purposes (Chen & Thissen, 1997), the data of student response on the Business English test used in this study adequately meet the assumption of local independence. Therefore, we have so far shown that the three major assumptions underlying IRT are met in the student response data that we analyzed in this study so that these results can provide support for the accuracy of the item and person parameter estimates that we obtained under the IRT model.

Item-Fit, Overall Model-Fit, and Item Parameter Estimates

Because the IRT is data-based, which means that the item and person parameters estimates are based on the observed response patterns of these students on the test items, we investigate to what extent the observed response patterns students on each item of the Business English test and on all of the test items across all students are consistent with the expected or predicted response patterns based on the model under investigation. Our first focus is on the consistency between the observed and predicted response patterns at the item level, where we tested the consistency using the signed chi-square (s-χ^2) statistic with a significance level of 0.05. An item is said to fit the model being investigated when the statistical value of the item is not statistically significant ($p \geq 0.05$). Through the item-fit test, we obtained the number of test items that fit each of the five models tested (see Table 2), where all items that fit the corresponding model have an RMSEA of less than 0.04. Table 2 demonstrates that of the five models tested, the 4PL model is the model with the most fit items. This indicates that the 4PL IRT model is better than the other four models for us to use to estimate item and person parameters.

Table 2. Summary of Item-Fit Analysis

Model Tested	The Number of Fit Item	%
Rasch	9	47.37
1PL	9	47.37
2PL	14	73.68
3PL	16	84.21
4PL	17	89.47

Our second focus is overall model-fit analysis to investigate a better model for estimating item and person parameters in terms of the data as a whole (i.e., response patterns across all items and all students analyzed in this study). We assessed overall model-fit based on a number of statistics and indices, one of which is M_2 statistic (Maydeu-Olivares & Joe, 2006). This statistic requires the model to adequately fit the overall data when the null hypothesis fails to be rejected. To determine which model is better among the models that fit well with the overall data, we choose the model with the smaller M_2 value. Table 3 demonstrates that the Rasch, 1PL, and 2PL models are statistically significant under the M_2 statistic, indicating that they have very poor fit with the responses to the overall data. We only failed to reject the null hypothesis on the 3PL and 4PL models, meaning that the two models fit the data well. By comparing the M_2 values of the two models, the 4PL model fits the overall data better than the 3PL model. We also carried out an overall model-fit analysis by considering RMSEA and SRMSR as fit indices by setting 0.08 as the cut-off value (Hooper et al., 2008). Based on the two fit indices, it turns out that all five models fit well with the overall data (Table 3). Because the model with smaller RMSEA and SRMSR values is better than another model in terms of its fit with the overall data, the 3PL model is the best in terms of RMSEA while the 4PL model is the best in terms of SRMSR.

We also used two other fit indices, namely AIC and BIC, where a model with a smaller value on that index means that the model is a better fit. Burnham and Anderson (Antoniou et al., 2022) suggested that when the difference in the estimated values of AIC and BIC is greater than 10 points, the difference can be considered as a meaningful difference with a model with smaller AIC and BIC values said to be a better-fitting model. The results of the analysis presented in Table 3 show that the 4PL model and 2PL model are the best-fitting models in terms of AIC and BIC indices, respectively. Lastly, we also examined the overall model-fit through a likelihood-ratio test with a log-likelihood with the consideration that a better-fitting model is one with a greater log-likelihood. Through this test, the 4PL model turns out to be the best-fitting model. Moreover, the 4PL model is also significantly the most fit compared to other simpler models in terms of the number of parameters involved (i.e., the 2PL model is significantly better fit than the 1PL model ($\chi^2(18) = 266.63$, $p < 0.001$), the 3PL model is significantly better fit than the 2PL model ($\chi^2(19) = 115.676$, $p < 0.001$), and the 4PL is significantly better fit than the 3PL model ($\chi^2(19) = 50.876$, $p < 0.001$). From a number of statistics and fit indices that we consider in the overall model-fit analysis, the 4PL model is the relatively best model in terms of its fitness with the overall data. Therefore, based on the results we obtained on item-fit and overall model-fit analyzes, the 4PL model is the best for estimating item and person parameters.

Table 3. Summary of Overall Model-Fit Analysis

Model Tested	Model-Fit Indices					
	M_2 (*df*)	RMSEA [95% CI]	SRMSR	AIC	BIC	logLik
Rasch	640.8411 (170) ***	0.053 [0.048, 0.057]	0.073	19794.31	19892.46	–9877.15
1PL	641.0403 (170) ***	0.053 [0.048, 0.057]	0.073	19794.31	19892.46	–9877.15
2PL	335.8679 (152) ***	0.035 [0.030, 0.040]	0.039	19563.68	**19750.17**	–9743.84
3PL	173.9199 (133)	**0.018 [0.009, 0.024]**	0.034	19486.00	19765.74	–9686.00
4PL	**157.4597 (114)**	0.020 [0.011, 0.027]	**0.030**	**19473.13**	19846.12	**–9660.56**

Note: RMSEA = Root mean square error of approximation, CI = Confidence interval, SRMSR = Standardized root mean square residual, AIC = Akaike information criterion, BIC = Bayesian information criterion, logLik = Log-likelihood, and *df* = the degrees of freedom. The values of the information criteria presented in bold font are indicative of a better model fit for a given criterion. *** $p < 0.001$

We have demonstrated that the 4PL model fits the response pattern on an item and the response pattern on all items and students so that the model is the best for estimating item and person parameters. Therefore, the properties of the items contained in the Business English test are based on four parameters, namely item difficulty (*b*), item discrimination (*a*), pseudo-guessing or lower asymptote (*g*), and carelessness or upper asymptote (*u*). Through the 'mirt' package, the four parameters were estimated using the fixed quadrature EM method (Chalmers, 2012, 2023) taking into account that this method is generally called effective and efficient under measurement condition that is unidimensional (Chalmers, 2012, 2023; Kalkan, 2022) as we found in our current study. The results of the estimation of the four item parameters are presented in Table 4. First, we focus on highlighting the item properties in terms of difficulty or location parameter. No definite criteria are found to categorize item difficulty levels under the IRT framework. However, it has been suggested that when the ability distribution is at a mean of 0 with a standard deviation of 1, the estimated parameter difficulty of an item should vary and be in the range from –2.00 to 2.00 (DeMars, 2010; Hambleton et al., 1991; Hambleton & Swaminathan, 1985). When the estimated value of the difficulty parameter of an item is close to –2.00, it means that the item is getting easier, while the estimated value is close to 2.00, it means the item is getting harder. We found several literatures (e.g., Adedoyin & Mokobi, 2013; Georgiev, 2008) that offer criteria for categorizing item difficulty on a logit scale under the IRT framework, in which the literature recommends items with medium difficulty when the estimated value of the difficulty parameter is in the range from –1.00 to 1.00. We then followed the guidelines proposed by Georgiev (2008) to categorize item difficulty other than medium, namely very easy ($b < -2.00$), easy ($-2.00 \leq b < -1.00$), hard ($1.00 < b \leq 2.00$), and very hard ($b > 2.00$). Using these guidelines, items in the Busi-

ness English test generally have a medium level of difficulty (M = –0.357, SD = 0.916), where we found four easy items, one very hard item, and the remaining medium difficulty items (see Table 4). Such distribution of items based on level of difficulty is in line with what DeMars (2010) suggested that items of medium difficulty should be used in large numbers on a test to more precisely measure the ability of the majority of students, although this can reduce the precision of measurement in students of extreme ability.

We further describe the item properties based on the discrimination or slope parameter (a). In order to make it easier for us to interpret the estimated value of the discrimination parameter for each item, we used the guidelines provided by Baker and Kim (2017). The strength of an item in differentiating students based on their latent trait level can be divided into five categories, namely very low ($0.01 \leq a \leq 0.34$), low ($0.35 \leq a \leq 0.64$), moderate ($0.65 \leq a \leq 1.34$), high ($1.35 \leq a \leq 1.69$), and very high ($a > 1.7$). Based on this categorization, we found that overall, the items contained in the Business English test have a very high power in discriminating students based on the latent trait (M = 2.661, SD = 1.212). The categories of discriminating power for the items on the test varied from moderate, high, to very high with the predominance of those in the very high category (84.21%). In fact, some items have an estimated value of the discrimination parameter that is out of the ordinary, where typically the estimated value for this parameter is in the range between 0 and 2 (Hambleton et al., 1991; Hambleton & Swaminathan, 1985).

Table 4. Summary Statistics for Item Parameter Estimates by Item Under the 4PL IRT Model

Item	a_i	b_i	g_i	u_i
Item 1	1.5556862	–1.6546697	0.0094401	0.9483183
Item 2	2.7148582	–0.7520136	0.2634997	0.9999546
Item 3	2.2961431	–0.3621948	0.3898563	0.9863581
Item 4	1.8198077	–0.9062742	0.0017970	0.9374254
Item 5	1.6820592	–1.2847370	0.0063288	0.9995147
Item 6	2.6613100	–0.1015988	0.2899071	0.9627361
Item 7	3.6269056	–0.6615458	0.5285124	0.9999742
Item 8	6.5070081	0.8048856	0.1104993	0.9038981
Item 9	2.4428593	–1.0326933	0.0445597	0.9441664
Item 11	2.3932332	–1.4091145	0.0355848	0.8583347
Item 12	1.9706128	–0.2421769	0.1751527	0.9976376
Item 13	2.5223933	–0.2121849	0.3618650	0.9997530
Item 14	3.7509019	0.3092440	0.2191204	0.8840679
Item 15	3.2930864	0.3306282	0.2281344	0.9995556
Item 16	2.2620852	2.3120939	0.1590696	0.9674843
Item 18	0.7862873	–0.6280422	0.0025212	0.4513170
Item 19	1.7619221	0.3186292	0.2122414	0.8322562
Item 20	2.8880256	–0.7003214	0.3304544	0.8576747
Item 21	3.6282244	–0.9110747	0.4479359	0.9026758

Note. a = Item discrimination parameter, b = Item difficulty parameter, g = Pseudo-guessing, guessing, or pseudo-chance-level parameter, and u = Carelessness, inattention, or slipping parameter

The property of the item is then reflected in the estimated value of the pseudo-guessing or lower asymptote parameter (g) which represents the probability that students with low ability can answer the item correctly because they make a guess. Intuitively, the probability for each option in the multiple-choice item to be selected is $1/k$, where k represents the number of options. Paying attention to this, it is clear that when the probability of students to answer a multiple-choice item correctly without proper knowledge or ability or it is purely based on random guessing behavior or with a certain guessing strategy is more than $1/k$, the multiple-choice item is considered to be not good. This has also been emphasized by Hulin, Drasgow, and Parsons (Retnawati, 2014) that the probability of a student answering an item correctly by guessing alone should be less than or equal to one over the number of options. Because the multiple-choice items contained in the Business English test use four options, i.e., one keyed option and three distractors, it

means that the estimate of the *g* parameter for each of these items should not be greater than 0.25. We have found that descriptively, all 19 tests we analyzed under the 4PL IRT model performed well in terms of estimation of the *g* parameter (M = 0.201, SD = 0.163). However, by looking at the *g* parameter estimates for each item, we found seven items (36.84%) whose *g* parameter estimates are more than 0.25 and there is even one item (i.e., item 7) whose estimated *g* parameter exceeds 0.5 (Table 4). This indicates that making random guesses or using certain guessing strategies on item 7 already provides greater assurance for students to arrive at the correct answer. The characteristic of such item is certainly not what the teacher or test developer expects because what is expected of them is that this item is able to reflect the real abilities of students or the responses that students give to this item are truly based on the thinking processes that students do by using the knowledge they have.

The last property of the test items that is also considered in this study is the upper asymptote or carelessness parameter (*u*). We have mentioned that the estimation of *u* parameter represents the probability of high ability students, those who are supposed to know the correct response to an item, give an incorrect response to an item. Although there is no consensus on the threshold to declare an item is good in terms of estimated *u* parameter, it has been suggested that the estimated parameter should remain high, at least 0.9 (Barnard-Brak et al., 2018). When following this suggestion, descriptively we can say that all items analyzed under the 4PL IRT model have been good (M = 0.918, SD = 0.126). However, if we take a closer look at the *u* parameter estimates for each item, there are five items (26.32%) whose *u* parameter estimates is below 0.9. A study by Loken and Rulison (2010) which explored the characteristics of self-report items that were intended to measure adolescent attitudes and behavior based on the 4PL IRT model also found that the estimated *u* parameter of these items is in the range of 0.76 and 0.89, which means that it is still below 0.9. In addition, in a study where one of the objectives was to investigate the characteristics of items used on the Mathematics test for eighth graders under the 3PL and 4PL IRT models, Doğruöz and Arikan (2020) found that the estimated value of the *u* parameter for the test items ranged from 0.52 to 1.00. Our study unfortunately yielded a more extreme result than what Doğruöz and Arikan (2020) had obtained, where we even found one item (i.e., item 18) where the estimated value of *u* parameter is below 0.5. When we review the meaning of the upper asymptote parameter which leads to careless behavior or feelings of anxiety that result in high ability students not being able to answer an item correctly, it is not easy for us to explain how it can be that the anomalous behavior or feelings are so severe so that it causes the probability of high ability students to answer correctly the item does not even reach 0.5. Furthermore, if we reconsider the fact that high ability students should be very careful or minimally careless in responding to an item, that student should have a greater probability of answering item 18 correctly ($u \geq 0.5$) than answering the item incorrectly ($u < 0.5$).

In addition to providing the estimated values of the four parameters attached to the English Business test items as presented in Table 4, we also provide a representation of the estimated values in the form of item characteristic curves (ICCs) (see Figure 6). The more difficult an item, the more to the right the position of the inflection point of the item's characteristic curve or the greater the value of theta (θ, student ability) required to be able to give the correct response to an item. For instance, by comparing the positions of the inflection points of the characteristic curves of item 5, item 11, and item 16, the inflection point of the characteristic curve of item 16 is on the rightmost position compared to the inflection point of the other two ICCs. This demonstrates that item 16 is the most difficult. Furthermore, if we compare the position of the inflection point of the ICC for item 16 with the position of the inflection point of the ICCs for the other 18 items, it turns out that the inflection point of the ICC for item 16 is the far right, so item 16 is the most difficult and this is supported by the estimated value of difficulty parameter as presented in Table 4. Next is the item discrimination parameter which is represented by the slope of the tangent at the inflection point of the ICC. The steeper the tangent line or the more contrasting the probabilities of two test takers whose abilities around *b* differ only slightly, the greater

the discriminating power of the item. Accordingly, Figure 6 clearly demonstrates that item 8 has the steepest tangent among the other, and thus item 8 is the greatest in terms of its power to discriminate the abilities of the examinees. The next two parameters, namely the lower asymptote and the upper asymptote, respectively represent the highest probability of low ability and high ability students to respond correctly to an item. It is clear that the highest lower asymptote is owned by ICC for item 7, which means that the highest probability of students with low ability to give the correct response by guessing can be achieved when they respond to item 7. Meanwhile, the lowest upper asymptote is owned by ICC for item 18, which represents that the maximum probability that students with high abilities may have to give the correct response to an item is lowest achieved when responding to item 18.

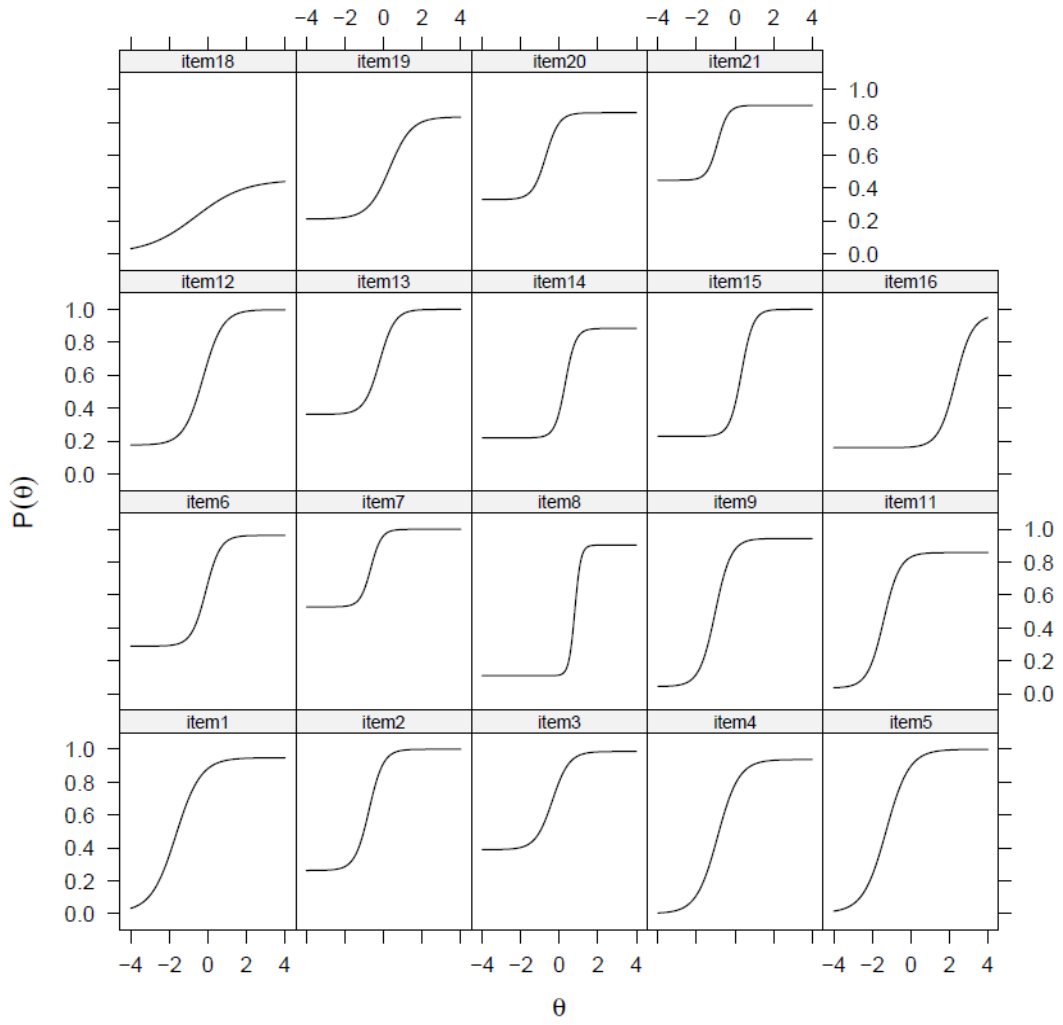


Figure 6. Item Characteristic Curves (ICCs) of 19 Items Analyzed in This Study Under 4PL IRT Model

In this study, as previously mentioned, we seek to provide a better understanding of the parameters of carelessness in the 4PL IRT model. Under this model, which allows the estimated value of the upper asymptote to be less than 1, high ability students who give the wrong answer to an item due to carelessness still have a high chance of answering the item correctly. The items contained in the Business English test which were analyzed based on IRT in this study in general have good characteristics in terms of the four parameters studied. However, there are two items that need more attention when our focus is on test development in terms of the estimated value of the pseudo-guessing parameter (g), namely item 7, and the carelessness parameter (u), namely item 18. In order for the low ability student's probability of answering item 7 correctly not to be

as high, i.e., it exceeds 0.5, as has been suggested by DeMars (2010), the distractors in item 7 should be revised so that they are more attractive and plausible to the low ability student. Meanwhile, as we mentioned earlier, we have difficulties in explaining the actual characteristics of item 18, how good or how bad the item is, in terms of the estimated value of the carelessness parameter. This is because if we revisit the meaning of the carelessness parameter, it is difficult for us to explain why high ability students are so careless that their probability of answering item 18 correctly is less than 0.5. In other words, we do not yet have strong reasons to judge the quality of item 18, apart from the high ability students' carelessness in responding to the item. Hence, the issue we found in item 18 remains a question for further studies. The issue we found in item 18 might be explained further by reviewing what has been suggested by Loken and Rulison (2010). By referring to the basic idea of introducing the upper asymptote parameter by Barton and Lord (1981), Loken and Rulison (2010) argued that this parameter reflects more on student behavior in responding to an item than the quality of the item itself, so that this parameter is more suitable to be a fixed characteristic of a test (i.e., the estimated value is constant for all test items) or an attribute of a person. Furthermore, according to them, determining the probability of a student with a certain ability to be able to give the correct response to an item not only considers the fluent response but also considers the non-faithful response, see Loken and Rulison (2010) for further explanation. The current study still considers the parameter of carelessness or upper asymptote as item-specific, in which the estimated value for each item varies.

Person-Fit and Person Parameter Estimate

Estimates of the item parameters that have been obtained were then used to obtain student ability or person parameter estimate. Before reporting further, the results of the person parameter estimate, we investigated how well a response pattern that a student gave to the 19 items contained in the Business English test matched the expected response pattern based on estimates according to the 4PL IRT model through person-fit analysis. The purpose of the analysis was to find out how accurate was the estimation of a student's ability according to the IRT model used, as well as to be an early detector of the existence of an unusual or a misfitting response pattern shown by a student and its possible causes. We conducted a person-fit analysis using the z_h statistic, a standardized fit index offered by Drasgow et al. (1985) whose values typically form a non-normal distribution (Felt et al., 2017). By using this statistic, a person misfit is detected when $|z_h| > 2$ (Desjardins & Bulut, 2018; Felt et al., 2017; Merino-Soto et al., 2023). Our analysis demonstrates that out of 1,000 students, there are a total of 24 (2.4%) students whose z_h value is less than –2 (none of the students whose z_h value is greater than 2), indicating that the response patterns on the 19 existing test items that they gave were inconsistent with the response patterns that the other 976 (97.6%) students gave as the expected response patterns according to the 4PL IRT model (Figure 7). As what has been conveyed by Felt et al. (2017), possible causes of the small number of misfitting response patterns here include guessing behavior, cheating behavior, lack of motivation, and students being distracted by something while responding to test items.

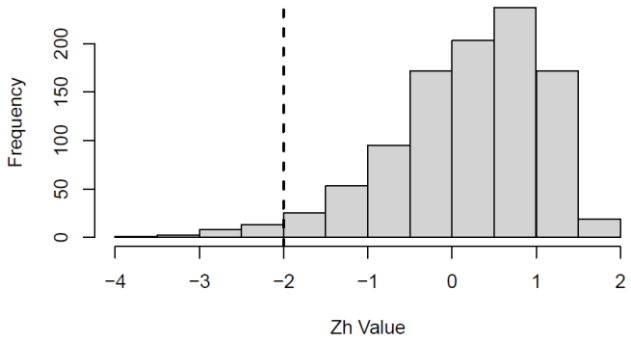


Figure 7. Distribution of z_h Values

Table 5. Summary of Person Parameter Estimate

Method	*M* (*SD*)	Min.	Max.	Skewness	Kurtosis	SE*
MLE	0.00 (0.91)	–2.18	1.71	–0.13	–0.67	0.41
MAP	–0.02 (0.83)	–2.04	1.44	–0.26	–0.73	0.36
EAP	0.00 (0.91)	–2.18	1.71	–0.13	–0.67	0.41

Note: MLE = Maximum likelihood estimation, MAP = Maximum a posteriori, EAP = Expected a posteriori, and SE* = Mean of standard error of person parameter estimate.

After presenting the results of the person-fit analysis, we present the results of the descriptive statistics of the person parameter estimate under the 4PL IRT model in Table 5. There were three methods that we used to estimate student ability, namely MLE, MAP, and EAP. Of the three methods, taking into account each value that is within two decimal places, it turns out that the MLE and MAP methods yielded the same results in descriptive statistics of person parameter estimate. The mean of estimated ability of the 1,000 students analyzed in this study is close to 0.00. Based on the mean of standard error of person parameter estimate, the estimated ability of students to respond to the 19 items contained in the Business English test obtained by the MAP method is the most accurate compared to that obtained by the other two methods. Studies conducted by Çetin and Çelikten (Doğruöz & Arikan, 2020) and Doğruöz and Arikan (2020) have shown that under 4PL IRT model, the MAP method provides the most accurate estimate of student ability than the EAP and MLE. Although a number of studies (e.g., Doğruöz & Arikan, 2020; Primi et al., 2018) have pointed out that there is no significant difference in the results of estimating students' abilities when estimated under the 4PL IRT model and under the simpler dichotomous IRT model (i.e., 2PL or 3PL), it is undeniable that the use of the 4PL IRT model can accommodate aberrant responses shown by a small number of high-ability test takers and increase the accuracy of person parameter estimation (Doğruöz & Arikan, 2020; Liao et al., 2012; Loken & Rulison, 2010; Primi et al., 2018; Rulison & Loken, 2009).

Measurement Precision

IRT allows us to identify how well each test item and the test provides information about the examinee's ability across different levels of latent trait (θ). This information basically reflects the level of precision of an item (represented by the item information function, IFF) or test (represented by the test information function, TIF) in revealing the ability of the examinee at a certain θ, where the higher the peak of an IFF or TIF, the higher the level of precision of the item or test in revealing the examinee's ability around θ when the information reaches a maximum. Several studies have revealed how to estimate this information under the 4PL model (see Liao et al., 2012; Loken & Rulison, 2010; Magis, 2013; Rulison & Loken, 2009; Waller & Feuerstahler, 2017) including how to estimate θ when the information reaches a maximum (see Magis, 2013). Items with higher information indicate that they produce more accurate ability estimates than items with lower information around a certain θ. Figure 8 presents the information function of the 19 items analyzed in this study under the 4PL IRT model. Figure 8 shows that of the 19 items, item 8 provides the highest maximum amount of information (i.e., around 6.6), indicating that item 8 yields the most accurate estimate of students' abilities with the highest accuracy achieved when the item is used for examining the ability of the student at around 0.8. Because the estimated value of the information of an item is inversely proportional to the square of the standard error of measurement (Hambleton et al., 1991; Hambleton & Swaminathan, 1985; Retnawati, 2014), it is logical to say that item 8 has the highest precision among other items in estimating students' abilities because it produces the smallest error. In contrast to item 8, item 18 has the lowest maximum amount of information (i.e., around 0.05) with the information function tending to be flat. This indicates that it is difficult for us to obtain accurate information about students' abilities through item 18. Considering that the estimated discriminating power parameter is directly proportional to the amount of information provided by an item (Hambleton & Swaminathan, 1985),

it is not surprising that the maximum amount of information provided by item 18 is so little because the estimated discriminating power parameter of the item is also low and it is compounded by the very low estimate of carelessness parameter (see Table 4).

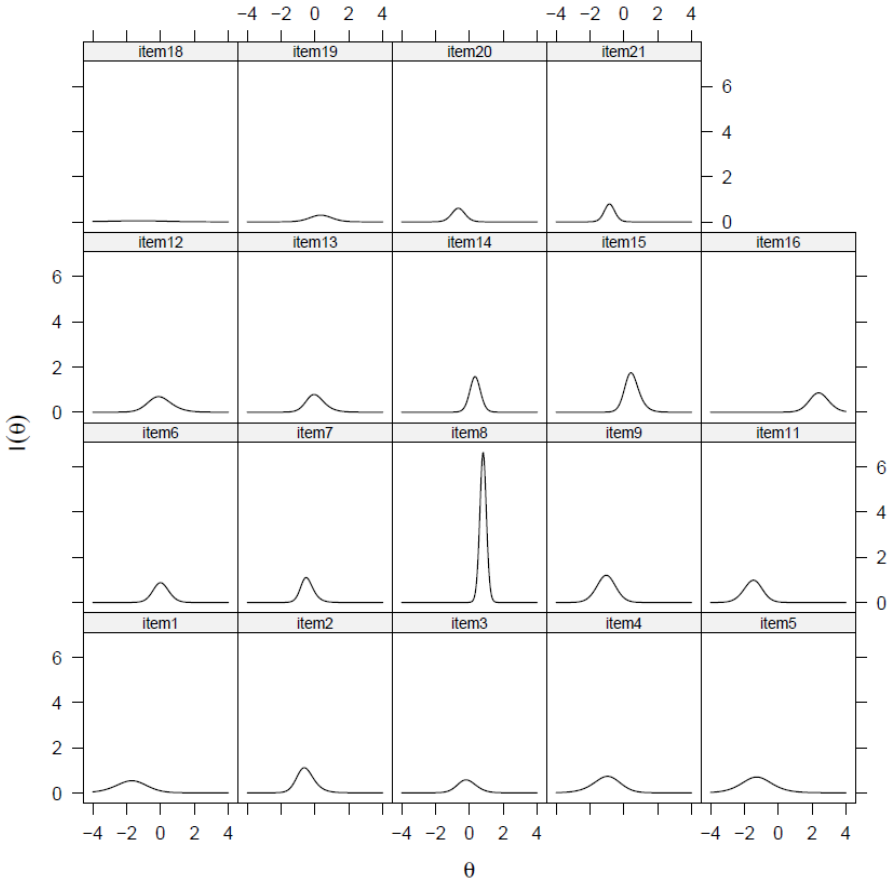


Figure 8. Item Information Functions (IFFs) of 19 Items Analyzed in this Study

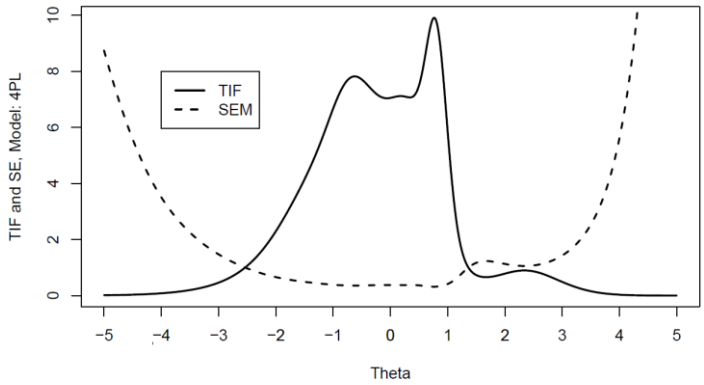


Figure 9. Test Information Function (TIF) and Standard Error of Measurement (SEM)

We have previously shown that the assumption of local independence as one of the three major assumptions of IRT has been satisfied. By satisfying this assumption, we can obtain the amount of information provided by the test, which is represented by TIF (see Figure 9), by adding up the amount of information provided by the 19 items (Hambleton et al., 1991; Hambleton & Swaminathan, 1985). With the same way of interpretation as IIF and by shifting the focus to a wider scope, namely the test, The Business English test which is composed of 19 items will provide accurate information about the level of ability of students when the test is administered to

those whose ability level is between about –2.6 to about 1.3. Furthermore, this test will provide the greatest amount of information, which means it will produce the most accurate estimate of student ability, when the test is administered to students whose ability level is around 0.8.

CONCLUSION

Our study seeks to provide additional understanding to the theory of the 4PL IRT model, particularly regarding the meaning and interpretation of the carelessness parameter as an additional parameter that establishes the model. The carelessness parameter indicates the probability that the examinee gives the correct response to an item with a condition that the correct response is based on sufficient knowledge so that the examinee does know the correct response. The meaning of the carelessness parameter clearly accommodates the aberrant responses from high ability examinees, thus allowing the estimation of the carelessness parameter to be less than 1. In ICC, the estimation of the carelessness parameter is represented by the height of the upper asymptote of the curve, which indicates the maximum possible probability that the high ability examinee can answer an item correctly. In order to provide further understanding of the 4PL IRT model including the meaning of the carelessness parameter, in this study we have presented a practical example of applying this model to identify the characteristics of several test items and test based on empirical data. Since the identification was carried out under the IRT framework, we have strived to present the practical example in detail starting from testing the three major assumptions of IRT, item-fit analysis, overall model-fit analysis, item parameter estimation, person-fit analysis, person parameter estimation, and estimation of measurement precision.

Although our study has attempted to provide an understanding of the 4PL IRT model and the carelessness parameter through a practical example using empirical data in detail, this study faced three limitations. The first limitation is that the estimation of item and person parameters was only based on several test items and several students who took the test, so the item and person parameter estimates do not fully reflect the actual conditions – this is indeed not the focus of our study. The second limitation is that this study considers the carelessness parameter as item-specific, while some studies suggest this parameter as a fixed characteristic of a test. The third limitation is that the item parameter estimates we obtained were only based on the EM method as the default in the 'mirt' package, while the 'mirt' package provides several methods based on the Bayesian framework. In light of the findings and limitations of this study, the 4PL IRT model is a promising model to use when the results of item-fit and model-fit analyzes support it considering that there is always the possibility that students with low or high abilities give an aberrant response. In addition, we expect further studies to compare ability estimates when the carelessness parameter is item-specific and it is a fixed characteristic of a test. Finally, further studies also can contribute to uncovering the characteristics of the item parameters based on methods other than EM.

ACKNOWLEDGMENT

We gratefully acknowledge financial support from *Lembaga Penelitian dan Pengabdian kepada Masyarakat* (LPPM, Institute of Research and Community Service), Universitas Terbuka through *Rencana Kerja dan Anggaran Tahunan* Universitas Terbuka (RKAT-UT), fiscal year 2023, which made this study possible to do and the results of this study possible to be published in an accredited national journal.

REFERENCES

Adedoyin, O. O., & Mokobi, T. (2013). Using IRT psychometric analysis in examining the quality of junior certificate mathematics multiple choice examination test items. *International Journal of Asian Social Science*, *3*(4), 992–1011. https://archive.aessweb.com/index.php/5007/article/view/2471

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

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, *42*(1), 7–16. https://doi.org/10.1097/01.mlr.0000103528.48582.7c

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

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), 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* (pp. 1–8) [Technical Report]. Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1981.tb01255.x

Battauz, M. (2020). Regularized estimation of the four-parameter logistic model. *Psych*, *2*(4), 269–278. https://doi.org/10.3390/psych2040020

Birnbaum, A. (1968). 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. 397–424). Addison-Wesley.

Bulut, O. (2015). Applying item response theory models to entrance examination for graduate studies: Practical issues and insights. *Journal of Measurement and Evaluation in Education and Psychology*, *6*(2), 313–330. https://doi.org/10.21031/epod.17523

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

Chalmers, R. P. (2023). *Package "mirt."* https://cran.r-project.org/web/packages/mirt/mirt.pdf

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. https://doi.org/10.3102/10769986022003265

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

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, *41*(3), 178–194. https://doi.org/10.1177/0146621616677520

DeMars, C. (2010). *Item response theory: Understanding statistics measurement*. Oxford University Press.

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

DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *The Canadian Journal for the Scholarship of Teaching and Learning*, *2*(2), 1–23. https://doi.org/10.5206/cjsotl-rcacea.2011.2.4

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

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*(1), 67–86. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Edwards, M. C., Houts, C. R., & Cai, L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological Methods*, *23*(1), 138–149. https://doi.org/10.1037/met0000121

Felt, J. M., Castaneda, R., Tiemensma, J., & Depaoli, S. (2017). Using person fit statistics to detect outliers in survey research. *Frontiers in Psychology*, *8*, 1–9. https://doi.org/10.3389/fpsyg.2017.00863

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). https://doi.org/10.5964/ejop.v4i3.431

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Haladyna, T. M., Rodriguez, M. C., & Stevens, C. (2019). Are multiple-choice items too fat? *Applied Measurement in Education*, *32*(4), 350–364. https://doi.org/10.1080/08957347.2019.1660348

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.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*(2), 139–164. https://doi.org/10.1177/014662168500900204

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*(1), 53–60. https://academic-publishing.org/index.php/ejbrm/article/view/1224

Houts, C. R., & Edwards, M. C. (2013). The performance of local dependence measures with psychological data. *Applied Psychological Measurement*, *37*(7), 541–562. https://doi.org/10.1177/0146621613491456

Kalkan, Ö. K. (2022). The comparison of estimation methods for the four-parameter logistic item response theory model. *Measurement: Interdisciplinary Research and Perspectives*, *20*(2), 73–90. https://doi.org/10.1080/15366367.2021.1897398

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

Kubinger, K. D., Holocher-Ertl, S., Reif, M., Hohensinn, C., & Frebort, M. (2010). On minimizing guessing effects on multiple-choice items: Superiority of a two solutions and three distractors item format to a one solution and five distractors item format. *International Journal of Selection and Assessment*, *18*(1), 111–115. https://doi.org/10.1111/j.1468-2389.2010.00493.x

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

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. *The International Journal of Educational and Psychological Assessment*, *1*(1), 1–11.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713–732. https://doi.org/10.1007/s11336-005-1295-9

Meijer, R. R., & Tendeiro, J. N. (2018). Unidimensional item response theory. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (Vol. 1, pp. 413–443). John Wiley & Sons. https://doi.org/10.1002/9781118489772.ch15

Merino-Soto, C., Angulo-Ramos, M., Rovira-Millán, L. V., & Rosario-Hernández, E. (2023). Psychometric properties of the generalized anxiety disorder-7 (GAD-7) in a sample of workers. *Frontiers in Psychiatry*, *14*, 1–16. https://doi.org/10.3389/fpsyt.2023.999242

Ogasawara, H. (2017). Identified and unidentified cases of the fixed-effects 3- and 4-parameter models in item response theory. *Behaviormetrika*, *44*(2), 405–423. https://doi.org/10.1007/s41237-017-0032-x

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

Paek, I., & Cole, K. (2020). *Using R for item response theory model applications*. Routledge.

Posit Team. (2023). *RStudio: Integrated development environment for R* (2023.6.0.421) [Computer software]. Posit Software, PBC. http://www.posit.co/

Primi, R., Nakano, T. D. C., & Wechsler, S. M. (2018). Using four-parameter item response theory to model human figure drawings. *Revista Avaliação Psicológica*, *17*(4), 473–483. https://doi.org/10.15689/ap.2018.1704.7.07

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), 1–15. https://doi.org/10.29333/pr/12657

Retnawati, H. (2014). *Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Nuha Medika.

Retnawati, H. (2016). *Analisis kuantitatif instrumen penelitian*. Parama Publishing.

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

Robitzsch, A. (2022). Four-parameter guessing model and related item response models. *Mathematical and Computational Applications*, *27*(6), 1–16. https://doi.org/10.3390/mca27060095

Rulison, K. L., & Loken, E. (2009). I've fallen and I can't get up: Can high ability students recover from early mistakes in CAT? *Applied Psychological Measurement*, *33*(2), 83–101. https://doi.org/10.1177/0146621608324023

Rupp, A. A., & Zumbo, B. D. (2004). A note on how to quantify and report whether IRT parameter invariance holds: When pearson correlations are not enough. *Educational and Psychological Measurement*, *64*(4), 588–599. https://doi.org/10.1177/0013164403261051

Rutkowski, L., von Davier, M., & Rutkowski, D. (Eds.). (2014). *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. CRC Press.

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

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia*, *126*(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864

Slocum-Gori, S. L., & Zumbo, B. D. (2011). Assessing the unidimensionality of psychological scales: Using multiple criteria from factor analysis. *Social Indicators Research*, *102*(3), 443–461. https://doi.org/10.1007/s11205-010-9682-8

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, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125–145. https://doi.org/10.1177/014662168400800201

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

Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexao e Critica*, *29*(1), 1–10. https://doi.org/10.1186/s41155-016-0040-x

Appendix 1. R Syntax for Test and Item Analysis Based on CTT Approach

```
#The initial data used in this analysis was saved in a file of type CSV (*.csv)
library(CTT)
#Import and read the data
data_initial <- read.csv2("dataforanalysisCTT.csv", sep = ";", header=T)
#Identify data dimension
dim(data_initial)
#Identify test and items characteristics
CTT_analysis <- itemAnalysis(data_initial)
CTT_analysis$itemReport
Summary_CTT_analysis <- rbind(CTT_analysis$nItem, CTT_analysis$nPerson, CTT_analysis$alpha,
                              CTT_analysis$scaleMean,CTT_analysis$scaleSD)
rownames(Summary_CTT_analysis) <- c('N of Item', 'N of Person', 'Alpha', 'Scale Mean',
                                    'Scale SD')
colnames(Summary_CTT_analysis) <- c(' ')
Summary_CTT_analysis
SEM <- CTT_analysis$scaleSD*sqrt(1-CTT_analysis$alpha)
SEM
#Save the results of test and item analysis based on CTT approach
sink('CTT Results.txt')
CTT_analysis$itemReport
Summary_CTT_analysis
cat('standard error of measurement (SEM):\n')
SEM
sink()
```

Appendix 2. R Syntax for Examining the Adequacy of IRT Assumptions

```
##Assumptions of IRT
library(mirt) #for examining parameter invariance
library(psych) #for examining unidimensionality
library(EFA.dimensions) #for examining local (in)dependence

data_initial <- read.csv2("dataforanalysis.csv", sep = ",", header=T)
datafull <- data_initial

##Assumption 1: Unidimensionality
cortest.bartlett(datafull)
KMO(datafull)
x_full <- cor(datafull)
#Generate scree plot and initial eigen value
y_full <- scree(x_full,factors = T, pc = F, main = "Scree plot", hline = NULL, add = F)
y_full_pc <- scree(x_full,factors = T, pc = T, hline = NULL, add = F)
initial_eigen_value <- data.frame(y_full$fv)
colnames(initial_eigen_value) <- "Eigen value"
initial_eigen_value
pca(datafull)
#Save scree plot
pdf(file = 'Scree Plot.pdf', paper = 'a4', height = 4, width = 6)
y_full <- scree(x_full,factors = T, pc = F, main = "Scree plot", hline = NULL, add = F)
y_full_pc <- scree(x_full,factors = T, pc = T, hline = NULL, add = F)
dev.off()

##Assumption 2: Parameter Invariance
##Assumption of (Item) Parameter Invariance
#Random splits to form subgroups of the same size (or relatively the same size)
set.seed(1)
sample <- sample(c(TRUE, FALSE), nrow(datafull),
                 replace=TRUE, prob=c(0.5, 0.5))
subgroup1 <- datafull[sample, ]
subgroup2 <- datafull[!sample, ]
```

```
#Item parameter invariance for person in subgroup 1
est_subgroup1_full <- mirt(data = subgroup1, model = 1, itemtype = "4PL", SE = TRUE,
SE.type = "sandwich", TOL = 0.001)
par_subgroup1_full <- coef(est_subgroup1_full, IRTpars = T, simplify = T)
#Item parameter invariance for person in subgroup 2
est_subgroup2_full <- mirt(data = subgroup2, model = 1, itemtype = "4PL", SE = TRUE,
SE.type = "sandwich", TOL = 0.001)
par_subgroup2_full <- coef(est_subgroup2_full, IRTpars = T, simplify = T)
#Generate data frame for item parameter for subgroup 1 and subgroup 2
x_full <- data.frame(par_subgroup1_full)
y_full <- data.frame(par_subgroup2_full)

#Correlation of a parameter for Subgroup 1 vs. Subgroup 2
(correlation_a_full <- cor.test(x_full$items.a, y_full$items.a, method = "pearson"))
correlation_a_full
#Save correlation of a parameter for Subgroup 1 vs. Subgroup 2
sink('Correlation of a parameter invariance.txt')
correlation_a_full
sink()
#Plot parameter invariance: a parameter
plot(x_full$items.a, y_full$items.a, main = "Scatter plot of a parameter",
     xlab = "Parameter 'a' in Subgroup 1", ylab = "Parameter 'a' in Subgroup 2",
     pch = 21, frame = FALSE)
abline(lm(y_full$items.a ~ x_full$items.a, data = datafull))
#Save plot parameter invariance: a parameter
pdf(file = 'Plot a parameter invariance.pdf', paper = 'A4', height = 4, width = 5)
plot(x_full$items.a, y_full$items.a, main = "Scatter plot of a parameter",
     xlab = "Parameter 'a' in Subgroup 1", ylab = "Parameter 'a' in Subgroup 2",
     pch = 21, frame = FALSE)
abline(lm(y_full$items.a ~ x_full$items.a, data = datafull))
dev.off()

#Correlation of b parameter for Subgroup 1 vs. Subgroup 2
(correlation_b_full <- cor.test(x_full$items.b, y_full$items.b, method = "pearson"))
correlation_b_full
#Save correlation of b parameter for Subgroup 1 vs. Subgroup 2
sink('Correlation of b parameter invariance.txt')
correlation_b_full
sink()
#Plot parameter invariance: b parameter
plot(x_full$items.b, y_full$items.b, main = "Scatter plot of b parameter",
     xlab = "Parameter 'b' in Subgroup 1", ylab = "Parameter 'b' in Subgroup 2",
     pch = 21, frame = FALSE)
abline(lm(y_full$items.b ~ x_full$items.b, data = datafull))
#Save plot parameter invariance: b parameter
pdf(file = 'Plot b parameter invariance.pdf', paper = 'A4', height = 4, width = 5)
plot(x_full$items.b, y_full$items.b, main = "Scatter plot of b parameter",
     xlab = "Parameter 'b' in Subgroup 1", ylab = "Parameter 'b' in Subgroup 2",
     pch = 21, frame = FALSE)
abline(lm(y_full$items.b ~ x_full$items.b, data = datafull))
dev.off()

#Correlation of g parameter for Subgroup 1 vs. Subgroup 2
(correlation_g_full <- cor.test(x_full$items.g, y_full$items.g, method = "pearson"))
correlation_g_full
#Save correlation of g parameter for Subgroup 1 vs. Subgroup 2
sink('Correlation of g parameter invariance.txt')
correlation_g_full
sink()
#Plot parameter invariance: g parameter
plot(x_full$items.g, y_full$items.g, main = "Scatter plot of g parameter",
     xlab = "Parameter 'g' in Subgroup 1", ylab = "Parameter 'g' in Subgroup 2",
     pch = 21, frame = FALSE)
abline(lm(y_full$items.g ~ x_full$items.g, data = datafull))
#Save plot parameter invariance: g parameter
pdf(file = 'Plot g parameter invariance.pdf', paper = 'A4', height = 4, width = 5)
```

```
plot(x_full$items.g, y_full$items.g, main = "Scatter plot of g parameter",
     xlab = "Parameter 'g' in Subgroup 1", ylab = "Parameter 'g' in Subgroup 2",
     pch = 21, frame = FALSE)
abline(lm(y_full$items.g ~ x_full$items.g, data = datafull))
dev.off()

#Correlation of u parameter for Subgroup 1 vs. Subgroup 2
(correlation_u_full <- cor.test(x_full$items.u, y_full$items.u, method = "pearson"))
correlation_u_full
#Save correlation of u parameter for Subgroup 1 vs. Subgroup 2
sink('Correlation of u parameter invariance.txt')
correlation_u_full
sink()
#Plot parameter invariance: u parameter
plot(x_full$items.u, y_full$items.u, main = "Scatter plot of u parameter",
     xlab = "Parameter 'u' in Subgroup 1", ylab = "Parameter 'u' in Subgroup 2",
     pch = 21, frame = FALSE)
abline(lm(y_full$items.u ~ x_full$items.u, data = datafull))
#Save plot parameter invariance: u parameter
pdf(file = 'Plot u parameter invariance.pdf', paper = 'A4', height = 4, width = 5)
plot(x_full$items.u, y_full$items.u, main = "Scatter plot of u parameter",
     xlab = "Parameter 'u' in Subgroup 1", ylab = "Parameter 'u' in Subgroup 2",
     pch = 21, frame = FALSE)
abline(lm(y_full$items.u ~ x_full$items.u, data = datafull))
dev.off()

##Assumption of (Person) Parameter Invariance
#Person parameter invariance
index_odd_full <- seq(1, dim(datafull)[2], 2)
index_even_full <- seq(2, dim(datafull)[2], 2)
item_odd_full <- datafull [ ,index_odd_full]
item_even_full <- datafull [ ,index_even_full]
#theta in odd order items
est_odd_full <- mirt(data = item_odd_full, model = 1, itemtype = "4PL", SE = TRUE,  SE.type
= "sandwich", TOL = 0.001)
theta_odd_full <- fscores(est_odd_full, method = "EAP", full.scores = T)
#theta in even order items
est_even_full <- mirt(data = item_even_full, model = 1, itemtype = "4PL", SE = TRUE,
SE.type = "sandwich", TOL = 0.001)
theta_even_full <- fscores(est_even_full, method = "EAP", full.scores = T)
#Generate plot person parameter invariance
x_full <- theta_odd_full[,1]
y_full <- theta_even_full[,1]
plot(x_full, y_full, main = "Scatter plot of ability",
     xlab = "Theta for items in odd order", ylab = "Theta for items in even order",
     pch = 21, frame = TRUE)
abline(lm(y_full ~ x_full, data = datafull))
#Correlation of theta in odd order items vs even order items
(correlation_theta_full <- cor.test(x_full, y_full, method = "pearson"))
correlation_theta_full
#Save correlation of person parameter invariance
sink('Correlation of person parameter invariance.txt')
correlation_theta_full
sink()
#Save plot person parameter invariance
pdf(file = 'Plot person parameter invariance.pdf', paper = 'A4', height = 4, width = 5)
plot(x_full, y_full, main = "Scatter plot of ability",
     xlab = "Theta for items in odd order", ylab = "Theta for items in even order",
     pch = 21, frame = TRUE)
abline(lm(y_full ~ x_full, data = datafull))
dev.off()

##Assumption 3: Local independence
options(max.print = 10000)
residu <- mirt(datafull, 1)
sink('Local dependence four statistics.txt')
```

```
cat('Yen Q3: n')
residuals(residu, type = "Q3", df.p = TRUE)#based on Yen's Q3
cat('LD Based on X2 Statistics: n')
residuals(residu, type = "LD", df.p = TRUE)#based on X2 statistics
cat('LDG2 Based on G2 Statistics: n')
residuals(residu, type = "LDG2", df.p = TRUE)#based on G2 statistics
cat('JSI Statistics: n')
residuals(residu, type = "JSI", df.p = TRUE)#based on Jack-knife slope index
sink()
```

Appendix 3. R Syntax for Estimating Item and Person Parameters Based on 4PL IRT Model

```
library(mirt)
library(psych)
#Input and read data that has no item whose point-biserial is non-positive
data_initial <- read.csv2("dataforanalysis.csv", sep = ",", header=T)
datafull <- data_initial
model1PL_full <- '
              F = 1-19
              CONSTRAIN = (1-19, a1)
            '
#STEP 1
#Data analysis with IRT approach: Rasch, 1PL, 2PL, 3PL, and 4PL
test_Rasch_full <- mirt(data = datafull, model = 1,  itemtype = "Rasch", SE = T, SE.type =
"sandwich", TOL = 0.001)
test_1PL_full <- mirt(data = datafull, model = model1PL_full, SE = T, SE.type = "sandwich",
TOL = 0.001)
test_2PL_full <- mirt(data = datafull, model = 1,  itemtype = "2PL", SE = T, SE.type =
"sandwich", TOL = 0.001)
test_3PL_full <- mirt(data = datafull, model = 1,  itemtype = "3PL", SE = T, SE.type =
"sandwich", TOL = 0.001)
test_4PL_full <- mirt(data = datafull, model = 1,  itemtype = "4PL", SE = T, SE.type =
"sandwich", TOL = 0.001)
#Item fit test based on chi-square
item_fit_Rasch_full <- itemfit(test_Rasch_full, fit_stats = "S_X2")
item_fit_1PL_full <- itemfit(test_1PL_full, fit_stats = "S_X2")
item_fit_2PL_full <- itemfit(test_2PL_full, fit_stats = "S_X2")
item_fit_3PL_full <- itemfit(test_3PL_full, fit_stats = "S_X2")
item_fit_4PL_full <- itemfit(test_4PL_full, fit_stats = "S_X2")
#Item fit test results based on chi-square
item_fit_Rasch_full$Note <- ifelse(item_fit_Rasch_full$p.S_X2 >= 0.05, 'Fit', 'Not Fit')
item_fit_1PL_full$Note <- ifelse(item_fit_1PL_full$p.S_X2 >= 0.05, 'Fit', 'Not Fit')
item_fit_2PL_full$Note <- ifelse(item_fit_2PL_full$p.S_X2 >= 0.05, 'Fit', 'Not Fit')
item_fit_3PL_full$Note <- ifelse(item_fit_3PL_full$p.S_X2 >= 0.05, 'Fit', 'Not Fit')
item_fit_4PL_full$Note <- ifelse(item_fit_4PL_full$p.S_X2 >= 0.05, 'Fit', 'Not Fit')
#Show the results of item fit test based on chi-square
item_fit_Rasch_full
item_fit_1PL_full
item_fit_2PL_full
item_fit_3PL_full
item_fit_4PL_full
#Summary of item fit test based on chi-square
summary_fit_full <- rbind(table(item_fit_Rasch_full$Note, exclude = 'Not Fit'),
table(item_fit_1PL_full$Note, exclude = 'Not Fit'),
                          table(item_fit_2PL_full$Note, exclude = 'Not Fit'),
table(item_fit_3PL_full$Note, exclude = 'Not Fit'),
                          table(item_fit_4PL_full$Note, exclude = 'Not Fit'))
rownames(summary_fit_full) <- c('Rasch', '1PL', '2PL', '3PL', '4PL')
#Show the summary of item fit test based on chi-square
summary_fit_full
#Model fit test based on AIC, BIC, Log-likelihood
anova(test_Rasch_full, test_1PL_full, test_2PL_full, test_3PL_full, test_4PL_full)
```

```
#Model fit test based on RMSEA (Note: copy and paste the result in your document as soon as
it appears in case you have a large sample size dan number of item)
M2(test_Rasch_full)
M2(test_1PL_full)
M2(test_2PL_full)
M2(test_3PL_full)
M2(test_4PL_full)
#Save the results of model fit test
sink('Results of Model Fit Test.txt', split = T)
cat('Model Fit Test Based on Chi-Square, Model: Rasch\n')
item_fit_Rasch_full
cat('\nModel Fit Test Based on Chi-Square, Model: 1PL\n')
item_fit_1PL_full
cat('\nModel Fit Test Based on Chi-Square, Model: 2PL\n')
item_fit_2PL_full
cat('\nModel Fit Test Based on Chi-Square, Model: 3PL\n')
item_fit_3PL_full
cat('\nModel Fit Test Based on Chi-Square, Model: 4PL\n')
item_fit_4PL_full
cat('\nSummary of the Results of Model Fit Test Based on Chi-Square:\n')
summary_fit_full
cat('\nResults of Model Fit Test Based on the AIC, BIC, Log-likelihood\n')
anova(test_Rasch_full, test_1PL_full, test_2PL_full, test_3PL_full, test_4PL_full)
cat('\nRMSEA, Model: Rasch\n')
M2(test_Rasch_full)
cat('\nRMSEA, Model: 1PL\n')
M2(test_1PL_full)
cat('\nRMSEA, Model: 2PL\n')
M2(test_2PL_full)
cat('\nRMSEA, Model: 3PL\n')
M2(test_3PL_full)
cat('\nRMSEA, Model: 4PL\n')
M2(test_4PL_full)
sink()
##STEP 2
#Item parameter estimation
itempara4PL_full <- coef(test_4PL_full, simplify=TRUE, IRTpars = T)
itempara4PL_full$items
#Save the results of item parameter estimation
sink('Results of Item Parameter Estimation 4PL Model.txt')
cat('\nItem Parameter Estimation 4PL Model\n')
itempara4PL_full$items
sink()
#ICC, IIF, TIF, and SEM Based on 4PL IRT Model
item_number <- ncol(datafull)
bwtheme <- standard.theme("pdf", color=FALSE)
plot(test_4PL_full, type = 'trace', theta_lim = c(-4, 4), which.items = 1:item_number,
     main = "ICC per Item", par.settings=bwtheme)
plot(test_4PL_full, type = 'infotrace', theta_lim = c(-4, 4), which.items = 1:item_number,
     main = "IIF per Item", par.settings=bwtheme)
#Save ICC dan IIF
pdf('ICC and IIF.pdf', paper = 'A4')
item_number <- ncol(datafull)
bwtheme <- standard.theme("pdf", color=FALSE)
plot(test_4PL_full, type = 'trace', theta_lim = c(-4, 4), which.items = 1:item_number,
     main = "ICC per Item", par.settings=bwtheme)
plot(test_4PL_full, type = 'infotrace', theta_lim = c(-4, 4), which.items = 1:item_number,
     main = "IIF per Item", par.settings=bwtheme)
dev.off()
#TIF and SE graph in the same scale
Theta <- matrix(seq(-5,5, 0.01))
tinfo <- testinfo(test_4PL_full, Theta)
theta_info <- data.frame(Theta, tinfo)
plot(Theta, tinfo, type = 'l', lwd = 2, ylab ="TIF and SE, Model: 4PL")
axis(side = 1, at = -5:5, tick = TRUE)
SE <- 1/sqrt(tinfo)
```

```
lines(Theta, SE, lty = 2, lwd =2)
legend(-4,8, legend = c("TIF", "SEM"), lwd =2, lty = c(1,2), cex = 1.0)
#Save TIF and SE graph in the same scale
pdf('TIF and SE graph in the same scale.pdf', paper = 'A4', height = 5, width = 10)
Theta <- matrix(seq(-5,5, 0.01))
tinfo <- testinfo(test_4PL_full, Theta)
theta_info <- data.frame(Theta, tinfo)
plot(Theta, tinfo, type = 'l', lwd = 2, ylab ="TIF and SE, Model: 4PL")
axis(side = 1, at = -5:5, tick = TRUE)
SE <- 1/sqrt(tinfo)
lines(Theta, SE, lty = 2, lwd =2)
legend(-4,8, legend = c("TIF", "SEM"), lwd =2, lty = c(1,2), cex = 1.0)
dev.off()
##STEP 3
#Person fit test based on 4PL IRT model
person_fit4PL <- personfit(test_4PL_full)
person_fit4PL
hist(person_fit4PL$Zh, xlab = "Zh Value")
abline(v = -2, lwd = 3, lty =2)
#Save person fit test results
sink('Results of Person Fit Test Based on 4PL IRT Model.txt')
cat('Results of Person Fit Test Based on Zh Value\n')
person_fit4PL
sink()
#Save histogram showing the results of person fit test
pdf(file = 'Person Fit Histogtam.pdf', paper = 'a4', height = 4, width = 6)
hist(person_fit4PL$Zh, xlab = "Zh Value")
abline(v = -2, lwd = 2, lty =2)
dev.off()
#Person misfit test
person_misfit4PL <- subset(person_fit4PL, Zh < -2)
person_misfit4PL
nrow(person_misfit4PL)
#Save person misfit test
sink('Results of Person Misfit Test Based on 4PL IRT Model.txt')
cat('Results of Person Misfit Test Based on Zh Value\n')
person_misfit4PL
nrow(person_misfit4PL)
sink()
##STEP 4
#Person parameter estimation (theta)
theta_MLE <- fscores(test_4PL_full, method = "ML", full.scores = T, full.scores.SE = T)
theta_MAP <- fscores(test_4PL_full, method = "MAP", full.scores = T, full.scores.SE = T)
theta_EAP <- fscores(test_4PL_full, method = "EAP", full.scores = T, full.scores.SE = T)
#Descriptive summary of person parameter estimation
describe(theta_MLE)
describe(theta_MAP)
describe(theta_EAP)
#Save person parameter estimation (theta)
sink('Results of Person Parameter Estimation 4PL Model.txt')
cat('Person Parameter Estimation 4PL Model\n')
theta_MLE
theta_MAP
theta_EAP
describe(theta_MLE)
describe(theta_MAP)
describe(theta_EAP)
sink()
##STEP 5
#Determine test empirical and marginal reliability based on 4PL IRT model
rxx_eap_full_4PL <- fscores(test_4PL_full, method = "EAP", full.scores.SE = TRUE)
empirical.reliability_full_4PL <- empirical_rxx(rxx_eap_full_4PL)
marginal.reliability_full_4PL <- marginal_rxx(test_4PL_full)
#Show empirical and marginal reliability based on 4PL IRT model
empirical.reliability_full_4PL
marginal.reliability_full_4PL
```

```
#Save empirical and marginal reliability based on 4PL IRT model
sink('Empirical and Marginal Reliability IRT.TXT', split = T)
cat('Empirical Reliability:\n')
empirical.reliability_full_4PL
cat('\nMarginal Reliability:\n')
marginal.reliability_full_4PL
sink()
#Plot Reliability 4PL
plot(test_4PL_full, type = 'rxx', theta_lim = c(-6, 6),
     main="Conditional Reliability_Full_4PL")
#Save conditional reliability plot
pdf('Conditional Reliability.pdf', paper = 'A4')
plot(test_4PL_full, type = 'rxx', theta_lim = c(-6, 6),
     main="Conditional Reliability, Model: 4PL")
dev.off()
```