Unpacking the Reading Subtest of ProTEFL: A 4PL IRT Model Study

Lovieanta Arriza [a,1,*], Mark Lester B. Garcia [b,2], Aulia Noor Rizarni [a,3]

[a] Universitas Negeri Yogyakarta, Jalan Colombo No.1, Yogyakarta 55281, Indonesia
[b] Ateneo de Manila University, Katipunan Avenue, Quezon City 1108, Philippines
[1] lovieanta0341pasca.2022@student.uny.ac.id; [2] mlgarcia@ateneo.edu; [3] aulianoor.2022@student.uny.ac.id
* Corresponding Author

ARTICLE INFO

Article history
Received August 22, 2025
Revised August 31, 2025
Accepted September 09, 2025

Keywords
Item Response Theory;
Four-Parameter Logistic Model;
Reading Comprehension;
ProTEFL

ABSTRACT

Reading comprehension is one of the key skills assessed in English language proficiency tests. However, many test items used in schools often fail to meet the expected psychometric standards, which can affect the accuracy of ability measurement. This study addresses the issue by evaluating the psychometric characteristics of the reading skill items in the ProTEFL instrument using an Item Response Theory (IRT) approach. This study contributes to the development of a high-quality standardized item bank for reading comprehension assessment and highlights the usefulness of the 4PL model for identifying problematic items. A quantitative descriptive method was employed through the analysis of 50 multiple-choice reading items from the ProTEFL. Responses from 8,038 test-takers were analyzed by checking sample adequacy, testing IRT assumptions, selecting model fits, and estimating item parameters. The results showed that the 4PL model provided the best fit (AIC = 468506.9; BIC = 469905.3; logLik = 234053.4), with 43 items meeting unidimensionality, local independence, and parameter invariance. About 70% of the items had good discrimination power, item difficulty was well distributed between 0 and 2 logits, pseudo-guessing values were mostly low, and 33% of items exhibited high upper asymptotes, indicating a risk of incorrect responses even by high-ability test-takers. These findings underscore the usefulness of the 4PL model in detecting flawed items and improving item quality. The study contributes to the refinement of the ProTEFL reading item bank and provides implications for developing valid and reliable language assessment instruments. Further research is recommended to apply multidimensional IRT models to cover other English language domains beyond reading.

1. Introduction

English is widely used in business, science, technology, and academic discourse, making it a necessary skill for professionals hoping to work globally. This has led to a transformation in English proficiency that has been a gate gateway to better job opportunities, higher salaries, and international mobility [1]. Beyond individual benefits, English has also become a global lingua franca that facilitates cross-cultural communication, knowledge exchange, and international collaboration, thereby reinforcing its role as a critical competence in the 21st century [2][3], which underscores the need for accurate and comprehensive assessments of English proficiency.

https://doi.org/10.59247/jtped.v2i2.27

English proficiency test is commonly conceptualized as a multidimensional construct that encompasses four modalities (reading, listening, speaking, and writing) [4]. Reading comprehension is an important part of language proficiency and one of the many requirements that must be fulfilled for each skill in the question package. Test questions can be developed as an efficient instrument for assessing students' reading proficiency based on a number of reading skills evaluation criteria. English proficiency test results are required by many universities as a condition of admission, as a requirement for exams, and as a requirement for graduation. Being able to communicate effectively in English is considered essential for networking and professional development, as well as a way to compete in a more worldwide labor market [5]. This development requires the measurement of language skills for students and other professionals.

There is a significant disparity between university graduates' actual English language proficiency and employers' expectations [6]. The challenges for measuring reading literacy skills include the influence of teachers' inexperience of the best way to formulate these skills, school facilities with regard to the availability of books and topics that are associated with reading, and factors dealing with students' starting proficiency from the very beginning [7]. Errors in item writing can make exam problems simpler or harder instead of what they might be. While certain errors give unprepared students hints that help them estimate the right response, unpleasant, overly complicated, or esoteric exam issues hinder prepared students from showcasing their comprehension of the material. In the context of the study conducted by [8], the two most frequent mistakes discovered are troublesome stems and implausible distractions.

The challenges presented in creating these reading questions highlight the necessity to analyze the quality and appropriateness of the materials used in reading literacy tests. The purpose of this assessment is to guarantee that the questions used accurately measure the required ability and not other unrelated factors (construct-irrelevant). When it comes to assessing language proficiency, mistakes in item drafting affect the test's validity and reliability in addition to its fairness.

Item Response Theory (IRT) is a popular method for doing an in-depth analysis of item quality. This theory offers advantages over Classical Test Theory (CTT) because it can independently estimate item parameters and participant ability, and provides richer information about the characteristics of each item [9]. IRT can also be applied in assessment systems that can automatically and accurately estimate ability scores [10]. Furthermore, IRT's capacity to distinguish between test difficulty and ability scores is one of its key benefits. Thus, ability estimates stay the same even when test takers' results may change based on the test's complexity [11].

IRT is a set of latent variable models widely used in the development of educational and psychological tests, such as standardized exams, personality tests, and certifications [12] and it also applied in international large-scale assessments (e.g., PISA, TIMSS) [13][14]. IRT is also frequently used to estimate student performance on college tests [15]. Specifically, in language assessments, such as reading, writing, listening, and speaking, IRT can help identify whether the test measures only one dimension or several dimensions simultaneously. For example, [16] used IRT to analyze the characteristics of English final exam items at SMA Negeri 5 Malang. By taking into account the characteristics of each item against the construct being measured, IRT allows for more precise and accurate ability measurement results, especially in identifying individual participant competencies [17].

Previous research has examined the psychometric characteristics of the ProTEFL listening section using IRT up to the 3PL model [18]. While these studies provide valuable insights into item functioning, they do not capture the upper asymptote effect, which accounts for the possibility of high-ability participants providing incorrect responses due to carelessness or other non-cognitive factors. To address this limitation, the present study focuses on the reading subtest and employs the 4PL model, thereby extending previous ProTEFL investigations and offering new contributions to the development of more reliable language assessment instruments.

Spesifically, this study aims to evaluate the quality of test items in the reading comprehension subdomain of the Proficiency Test of English as a Foreign Language (ProTEFL), an English proficiency examination developed by the Language Development Center of Yogyakarta State University (UNY). Considered equivalent to the TOEFL, ProTEFL scores are widely used to predict students' TOEFL performance and serve as a graduation requirement at UNY [19]. The test assesses three subdomains of English: listening comprehension, structure and written expression, and reading comprehension. Each section consists of 50 multiple-choice questions, which must be completed within two hours [20].

Given its function as a high-stakes test for graduation, the quality of ProTEFL items is critical. However, test items, particularly in the reading comprehension subdomain, may vary in quality due to challenges in item-writing practices, alignment with intended constructs, or contextual factors in language learning. Conducting item analysis using IRT provides an opportunity to examine item parameters more rigorously and to refine the ProTEFL item bank. This analysis is both urgent and timely to ensure that ProTEFL maintains validity and reliability comparable to other standardized English proficiency test. Prior research, such as the development of the University Reading and Comprehension Strategies Assessment (URCSA), showed that IRT analysis provides strong evidence of reliability, internal validity, and sensitivity in identifying strengths and weaknesses in students' reading skills [21].

The findings of this study are expected to support the development of a more valid, reliable, and representative assessment of reading comprehension. Moreover, refining the ProTEFL item bank benefits multiple stakeholders: students, by ensuring fairer and more accurate evaluations of their ability; test developers, by providing empirical evidence for item revision; and institutions, by upholding the credibility of ProTEFL as a standardized assessment instrument.

2. Method

This research is an exploratory descriptive study with a quantitative approach aimed to evaluate the psychometric characteristics of reading ability items. This type of research is widely adopted and valuable, providing data summaries and interpretations [22]. The main focus of this study is to estimate item parameters using an IRT model and analyze the quality of the reading ability instrument based on these estimation results. The overall procedure of this research followed several sequential stages, as illustrated in Fig. 1.

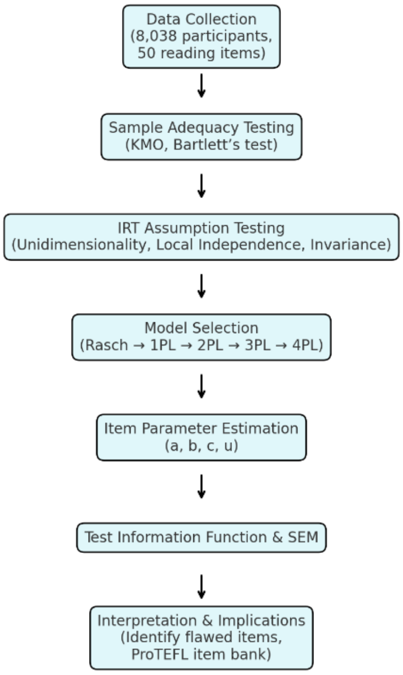

Fig. 1. Research flowchart

2.1. Subject and Data

The data were obtained from responses of 8.038 ProTEFL test-takers, consisting of undergraduate students from various faculties at UNY who took the test as part of their graduation requirement. This instrument is commonly used to measure English language skills of students and lecturers on college. Each participant completed 50 multiple-choice items in the reading comprehension section. The reading comprehension questions consist of short to medium-level texts followed by literal and inferential comprehension questions.

2.2. Proposed Method

Data analysis was conducted through several systematic stages to ensure that the data met the requirements of the IRT approach and to determine the best IRT model suited to the data characteristics. To ensure the validity of the results, all stages of the analysis were conducted using RStudio, supported by a specialized analysis package such as “psych” and "mirt" [12].

- Sample adequacy was checked using the Kaiser-Meyer-Olkin (KMO-MSA) test and data homogeneity using Bartlett's test. If the KMO-MSA value is >0.5 and the Bartlett's significance test is <0.05, then the data meets the requirements for further IRT assumption testing [23].
- IRT assumption testing needs to be met before selecting an IRT model. The assumption of unidimensionality of the data, meaning all items measure the same construct, is tested using factor analysis (FA) or principal component analysis (PCA) to ensure the dominance of one main factor [24][25]. The assumption of local independence, which requires participants' responses to one item to be independent of other items, can be considered if the unidimensionality assumption is met [26]. The invariance assumption is verified by estimating item parameters across different groups of test takers. The estimation results are presented in the form of a scatterplot. If the points on the scatterplot are close to the line passing through the origin with a slope of 1, or indicate highly correlated items, then the parameters are considered invariant [27][28].
- IRT model selection was performed by comparing the fit between five IRT models: Rasch, 1-Parameter Logistic (1PL), 2PL, 3PL, and 4PL. The selection of the best model was based on three main indicators: the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and the log-likelihood (logLik) [29]. Furthermore, the evaluation of the significance of the improvement in fit between nested models was reviewed using the Likelihood Ratio Test (LRT) [30]. This test produces a standardized chi-squared statistic ($S\chi^2$) [31], degrees of freedom (df), and a significance value (p-value). The model with the lowest AIC and BIC, the highest log-likelihood, and a significant improvement in fit based on the LRT was selected as the best model for parameter estimation.
- Item parameter estimation. The best model was then used to estimate the item parameters of the reading ability test. In this study, the 4-PL model [32] was selected as the most appropriate model, which estimates four item parameters: (a) discrimination power, (b) item difficulty, (c) pseudo-guessing, and (u) upper asymptote. This makes the 4PL particularly suitable for analyzing reading comprehension items, where misinterpretation of text or distraction can occur even among proficient test-takers. The criteria for good item parameters are presented in Table 1.

Table 1. Item parameter criteria in the 4PL IRT model

Parameter	Thresholds
a	Range from 0 to 2 [33]
b	Range from -2 to 2 [34]
c	Smaller values (approaching 0) [26]
u	Does not deviate significantly from other u-parameters [35]

- Evaluation of the Test Information Function (TIF) and Standard Error of Measurement (SEM) was carried out as a final stage to see how much and how accurate the information provided by all tests is in the range of participant abilities [36].

3. Results and Discussion

3.1. Data Feasibility Testing

Before testing the IRT assumptions, data suitability was first tested using KMO-MSA and Bartlett's Test of Sphericity. These analyses were conducted using the *KMO()* and *cortest.bartlett()* functions from the psych package in R. The test results are presented in Table 2. Based on Table 2, the feasibility test results show a KMO value of 0.9, which meets the requirement of > 0.5, and the Bartlett's test results with a p-value = 0.000 (p < 0.05), indicating that the response data of 8,038 test participants who answered the reading ability questions showed a sufficient inter-item correlation and met the elements of sample feasibility and data homogeneity [37][38]. This is further reinforced by [39][40] that the higher the KMO-MSA value and the more significant the Bartlett's [41], the better the factor analysis determination. Furthermore, this proof of feasibility provides a strong basis for conducting IRT assumption test analysis [42].

Table 2. Sample feasibility test results

Testing	Result
KMO − MSA	0.9

Bartlett′s	Khi-suqared = 25255
	Degree of freedom = 1225
	p-value = 0.000 (< 2.2e-16)

3.2. IRT Assumption Testing

Testing the IRT assumptions begins with a unidimensional analysis of the data. Participant response data must have one dominant component, indicated by the largest eigenvalue. A common rule of thumb is that the eigenvalue of the first contrast (i.e., the second component) should be less than 2.0, indicating that the residual variance does not form a meaningful secondary dimension [43]. This criterion supports the interpretation that the test items collectively measure a single dominant latent trait. This test is performed using PCA on the residual matrix [44][45]. The eigenvalue results from the PCA with RStudio are shown in Table 3. Table 3 shows that participants' responses to the reading ability questions yielded six components, with one dominant component having an eigenvalue of 4.59. The second and subsequent components ranged from 1.09 to 1.54. The ratio between the first and second components was 3.05, indicating that the first component was the most dominant. This is also supported by the cumulative percentage of the six components, which is 25.52%, which aligns with the concept proposed by [28] that the IRT unidimensionality assumption is considered met if the cumulative percentage of the components exceeds 20%. In addition, the proof of the assumption of data unidimensionality was also carried out using the factor analysis method and the results were presented in the form of a scree plot (Fig. 2).

Examining the scree plot in Fig. 1 reveals only one dominant factor in the reading ability test (eigenvalue greater than 4). Furthermore, the other two factors also contribute small eigenvalues to the explained variance component. Fig. 1 also shows that the eigenvalues begin to plateau from the fourth factor onward. These results show that the response data for reading ability questions only measures one factor, which is then referred to as a dimension. The results of the scree plot of factor analysis have provided visual information that the response data of the reading ability test participants used in this study have one dominant component [46]. This analysis aligns with [47], who emphasized the importance of understanding the dimensionality of data from language ability measurements in the IRT approach.

After the unidimensional assumption is met, the next step is to test the local independence assumption. In its proof, local independence can be met if the participant's answer to one item does not affect their answer to another item [48]. However, the practice in this study refers to [26], who stated that when the unidimensional assumption is met, local independence is also obtained. This is based on the concept that local independence is obtained from a predetermined complete latent space, namely when all ability dimensions that influence test performance have been taken into account. Furthermore, we tested the parameter invariance assumption, a characteristic assumption in IRT that differentiates and addresses the weaknesses of classical test theory [27]. The parameter invariance assumption was tested by estimating the parameters on different groups of test takers, in this case, categorized as odd-even. The results of this estimation were then presented in the form of a scatter plot, as shown in Fig. 3.

The scatter diagram points in Fig. 3 approach/follow the straight-line pattern $y = x$. The line $y = x$ is indicated by a red straight-line curve that passes through the origin (0,0) and has a gradient = 1. This result indicates that the reading ability test items have fulfilled the invariant assumption element [28]. In addition to item parameters, the invariance of ability parameters also needs to be proven. In the same way, the invariance of ability parameters is proven by classifying the item division, in this study the items are categorized into odd-even and the results of the invariance assumption can be seen in Fig. 4. To strengthen the evidence, this study also conducted a correlation analysis as another method for reviewing parameter invariance. The results of the positive correlation of each parameter tested are presented in Table 4. Referring to the three previously proven IRT assumptions, the information obtained indicates that the response data from 8,038 reading ability test participants met the IRT assumptions. Therefore, it can be concluded that the data used is suitable and ready for analysis to assess the quality of the instrument using the IRT approach.

Table 3. PCA results

Component	Eigenvalue	Variance percent	Cumulative variant percent
1	4.59587	11.068806	11.06881
2	1.546353	3.596169	14.66497
3	1.310134	3.046824	17.71180
4	1.147792	2.669283	20.38108
5	1.115687	2.594621	22.97570
6	1.093863	2.543867	25.51957

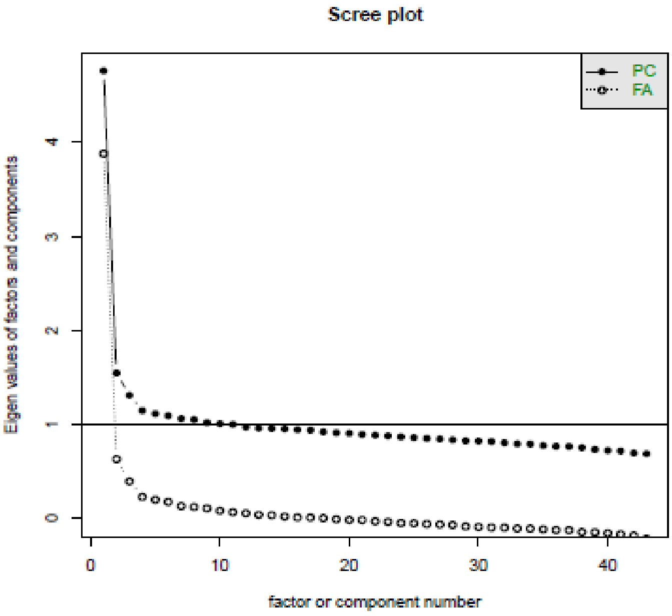

Fig. 2. Scree plot of FA and PCA results

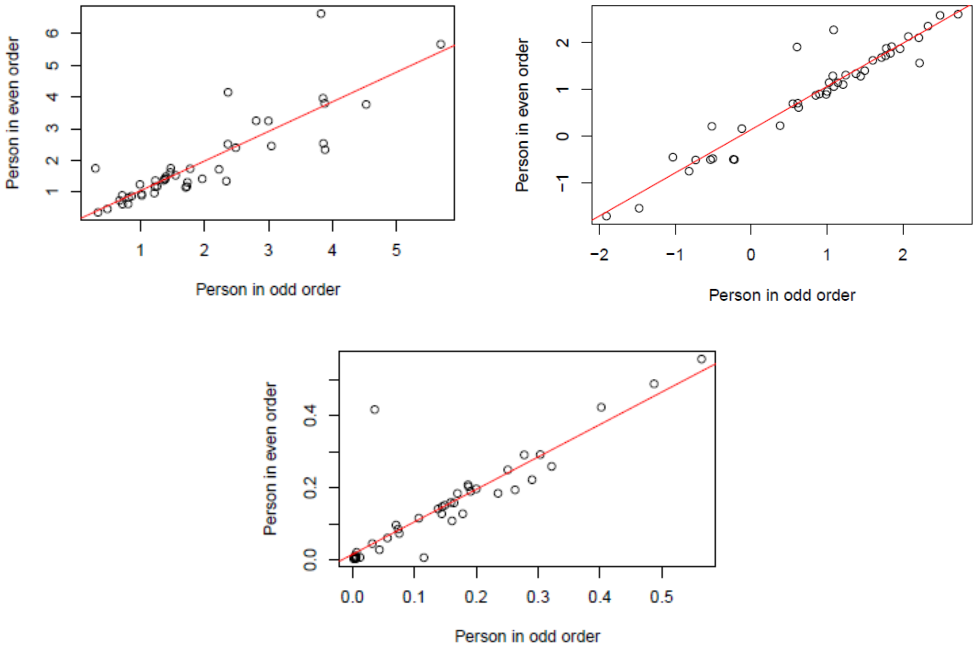

Fig. 3. Item parameter invariance results *(a, b, c)*

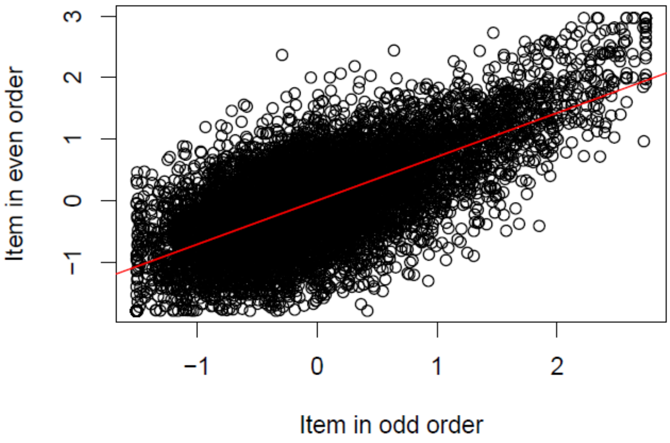

Fig. 4. Results of invariance of participant ability parameters

Table 4. Parameter invariance correlation results

Parameter	Correlation
Discrimination power *(a)*	0.855
Item difficulty (*b*)	0.954
Pseudo-guessing (*c*)	0.884
Ability	0.696

3.3. IRT Model Fit Test

After the IRT assumptions are met, the analysis using the IRT approach begins by testing the fit of the data to five tiered IRT models, namely Rasch, 1PL, 2PL, 3PL, and 4PL [49]. Table 5 shows the results of the suitability test of the reading ability instrument model against the five models which are the basis for selecting the best model for estimating the item parameters. Based on Table 5, it was discovered that the 4PL model performed the best at capturing the data, as seen by its lowest AIC value (468506.9), slightly higher BIC (469905.3), and highest log-likelihood (-234053.4). The LRT test [52], which demonstrated that each additional parameter from a simpler model to a more complicated model led to a statistically significant increase in model fit, further corroborated these fit results. The chi-square value (χ^2) in the 4PL model was 104.677 with df = 50 and p < 0.001, indicating that although the increase in fit from 3PL to 4PL was relatively small compared to the difference in the previous model, the increase was still statistically significant. Meanwhile, the comparison test between the Rasch and 1PL models is not shown further because both have identical logLik values, so the difference cannot be tested statistically (df = 0; χ^2 is invalid).

While global model fitting (AIC/BIC/logLik) is important, it only indicates overall model fit. The item fit method, on the other hand, provides information on how well each item fits the IRT model. Additional evaluation was conducted on the item fit of each model using standardized chi-square (Sχ^2), as was done by [53]. The results showed that the 4PL model produced the highest number of fit items compared to the other models, namely 43 out of 50 items. Based on this, the 4PL model was selected as the best model for item parameter estimation in this study. This also takes into account the characteristics of large response data [54][55], which is sufficient for use in advanced IRT models with stability and accuracy. The 4PL model is an extension of the 3PL model [56], which assumes that high-ability participants can still answer incorrectly due to factors such as carelessness, as indicated by an upper asymptote parameter that is not equal to one [35]. The choice of this model is an advantage of IRT because it allows estimation of participant ability from any combination of available items, provided the selected model fits the data [11].

Table 5. Summary of data fit comparison against five IRT models

Model	AIC	BIC	logLik	χ^2	df	p-value	Item count
Rasch [50]	475276.2	475632.8	-237587.1	-	-	-	5
1PL [51]	475276.3	475632.9	-237587.1	-0.103	0	NaN	5
2PL [51]	470729.6	471428.8	-235264.8	4644.709	49	<0.001	21
3PL [34]	468511.6	469560.4	-234105.8	2318.021	50	<0.001	42
4PL [32]	468506.9	469905.3	-234053.4	104.677	50	<0.001	43

3.4. Distribution of Item Parameters

After the IRT model was determined, an analysis was conducted to review the characteristics of the reading ability items. Item characteristics were described from 43 items that met the fit criteria based on the $S\chi^2$ $(p > 0.05)$. Therefore, Fig. 5 presents the distribution of the estimated results for each item parameter, focusing on the 43 items that fit the 4PL model. Based on Fig. 5, 13 of the 43 items (30%) had an a-parameter of more than 2, thus failing to meet the ideal criteria. An a-parameter hat too high suggests that the item is too sharp and over-discriminating, which risks making the estimation of participants' abilities unstable or focusing too much on small differences in participants' abilities. Meanwhile, the majority of the remaining items (70%) fell within the ideal discriminatory power range, with the highest concentration at a-parameter of 1.0 to 1.5. These items demonstrate fairly good and realistic discriminatory power [33] in the context of a reading ability test. Although most items have good discriminatory power, the presence of items with high a-parameter indicates a potential imbalance in the test characteristics, and some of these items require substantial review to maintain the quality of the instrument and measurement results.

Furthermore, for the item difficulty level parameter (b), Fig. 5 shows that the distribution of parameter estimates is quite spread out, concentrated in the range of 0 to 2. However, a more detailed analysis reveals that there are 6 items (around 14%) that have a b-parameter value greater than 2. This indicates that these items tend to be too difficult for the majority of participants. This can reduce the effectiveness of the items in measuring participants' abilities evenly, especially if the majority of participants are at a medium level of ability. Meanwhile, no items were found with a b-parameter value less than -2. This indicates that no items are too easy. The fairly wide distribution of b-parameter and the dominance of medium to high levels of difficulty illustrate that this reading ability instrument is more accommodating to measuring participants with intermediate to high abilities.

The distribution of the pseudo-guessing parameter (c) as shown in Fig. 5 shows that most items have very low c-parameter, close to 0. This is a good sign because it indicates that the probability of participants answering correctly randomly (guessing) is very low, which ultimately increases the validity of measuring the participants' true ability. The peak of the distribution is also very close to 0, which illustrates the effectiveness of the items in minimizing the guessing factor. However, there are several items with c-parameter above the acceptable limit for items with 5 answer options, namely 0.2. These items require special attention because they are indicated as items that are too easy to guess, and may not be effective in differentiating participants' abilities because luck plays a large role. These items should be re-evaluated, possibly revised or removed.

Meanwhile, for the distribution of the upper-asymptote parameter (u) in Fig. 5, it can be seen that this parameter indicates the maximum probability of a participant answering an item correctly, even if the participant has high ability. In this analysis, a minimum limit of 0.9 was used as a practical reference. The distribution results show that the majority of items (67%) have a u value ≥ 0.9, which indicates that the probability of participants with high ability answering correctly is close to perfect. However, there are 14 items (33%) that have a u-parameter below the 0.9 limit. This indicates that even though participants have high ability, there is still a possibility of not being able to answer the item correctly. This possibility could occur due to factors such as carelessness, question ambiguity, unrepresentative item content, or other non-cognitive factors. This is in line with the findings of [35] that the 4PL model can help identify flawed items due to construct-irrelevant variance. Revising or replacing such items is crucial to ensure that the ProTEFL reading subtest accurately measure proficiency rather than test-taking strategies or carelessness. However, in this study the distribution of the u-parameters still shows a positive trend because there are no items with extremely low u-parameter. The u-parameter are generally still within the acceptable range for academic measurement.

After estimating the item parameters, the next step is to evaluate the TIF. The TIF analysis aims to determine the extent to which the reading ability instrument provides accurate information in measuring test takers' abilities at various ability levels. Using 43 items that have been found to fit the 4PL model, the TIF was analyzed to determine which ability ranges were measured with high accuracy and to what extent the test was effective in differentiating participants at specific ability levels. In addition to the TIF, an SEM graph is also included which provides an overview of the magnitude of measurement error at each ability level. SEM has the best relationship with TIF. Ideally, when the test information is high, the SEM is lower, indicating a more accurate measurement. The following graphs present the TIF and SEM results.

The blue curve in Fig. 6 is a TIF graph showing the maximum information from the reading ability instrument, which is in the theta (θ) range of around +2 logits, where the information value is close to 20. This indicates that this reading ability test has high measurement accuracy for participants with high abilities, especially at above-average levels. Conversely, the measurement information decreases sharply in the $\theta < 0$ range, meaning this test is less accurate in measuring participants with low abilities. This is supported by the SEM curve which shows the lowest point of measurement error of < 1 at θ around +2 logits. In addition, the TIF and SEM curves can be seen to intersect visually at around θ = -2 and θ = +4. These points indicate the limits of the instrument's effectiveness, confirming that the optimal range of abilities that can be measured by this reading ability instrument lies between these two points. This is in line with [57] who stated that the theta (θ) range generated from the TIF and SEM curves indicates the extent of the test participant's abilities that can be measured by the instrument. It is important to note that the abilities of all test participants are not at the same level [58] [59].

This TIF pattern supports the information generated from item parameter estimation. The item difficulty ranges from 0 to +2 (mid-to-high proficiency), the low probability of guessing items, and the upper-asymptote parameter, indicating that high-ability participants could only actually answer a few difficult items almost perfectly, are further reinforced by the TIF graph, which reflects that the reading ability instrument was designed to provide accurate measurements for test takers with mid-to-high proficiency, rather than for those with low proficiency. Given that this test is used in the context of selection or meeting specific requirements (e.g., graduation or program acceptance), the dominant measurement focus on the positive θ range is considered relevant. Most items fall within the positive θ range, causing the test to be unbalanced in measuring low proficiency. The exclusion of the low θ range is not a weakness; rather, it further demonstrates that this instrument was designed to distinguish participants with mid-to-high proficiency, in line with the academic function of the English proficiency test. Therefore, this reading ability instrument is indeed more suitable for use in the context of selection or certification rather than for initial diagnostic purposes.

The findings of this study indicate that the 4PL model is capable of capturing variation not covered by the 1PL to 3PL models, namely identifying and eliminating items susceptible to guessing behavior [29]. The reason for using this model is not only due to statistical fit, but also because the 4PL model is successfully identified careless responding and items with construct-irrelevant variance in cognitive assessments [17]. The superior fit of the 4PL model in this study highlights its potential to advance language assessment research in Indonesia, where most item analyses have been limited to Rasch, 2PL, or 3PL models. Thus, the measurement will provide more accurate results, and the resulting scores will truly represent the test taker's reading ability more fully, compared to using only the simpler IRT model.

Responses that do not reflect true ability, such as correct answers due to guesswork and incorrect answers due to carelessness and inattention, pose a significant threat to the validity of test data and psychometric measurements because they can lead to erroneous ability estimates [60][61]. This study addresses [62] suggestion to empirically test the superiority of the 4PL model. The results show that the 4PL model considers the possibility of incorrect answers in high-ability participants, resulting in more accurate ability estimates. This is particularly relevant considering that a previous study by [63] showed that careless responding is a common problem in large-scale tests and can lead to distorted results if not modeled appropriately. Utilizing the 4PL model can provide more accurate ability estimates in the context of mass measurement tests. Thus, the 4PL model can provide new insights into the characteristics of the measurement instruments used [64]. The potential application of the 4PL model is wide open in the development of adaptive tests and more precise ability measurements due to its flexibility in modeling the upper limit of the probability of participant success.

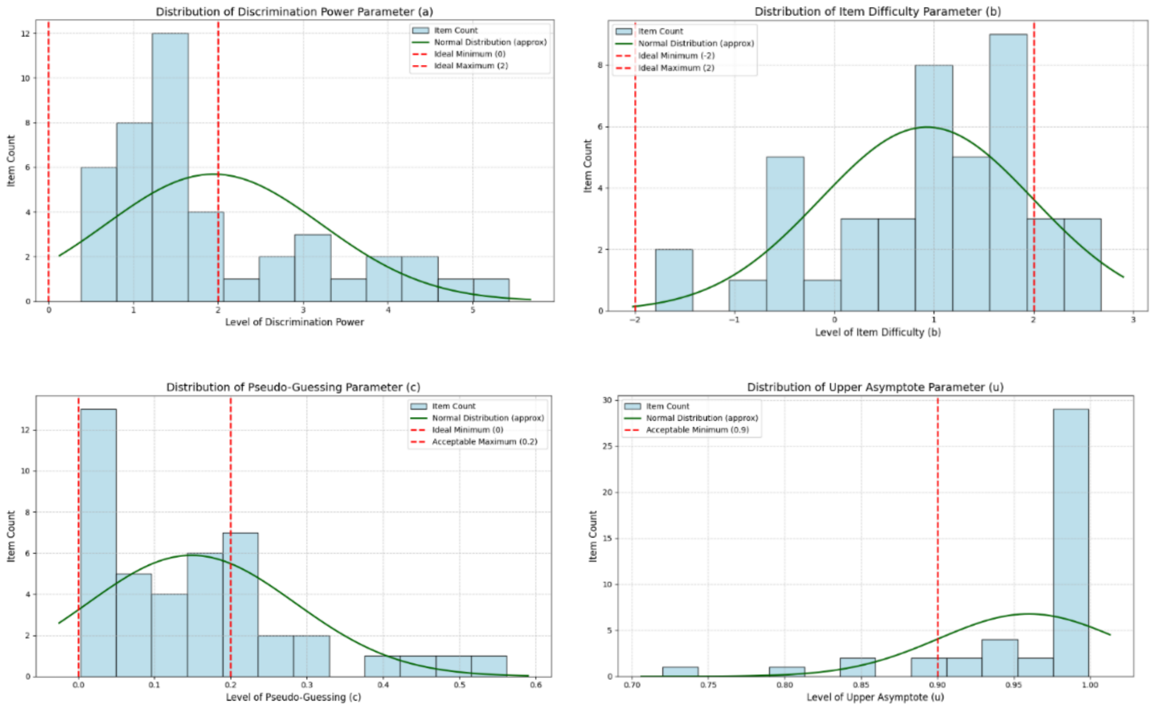

Fig. 5. Distribution of estimated a, b, c, and u parameters based on the 4PL model

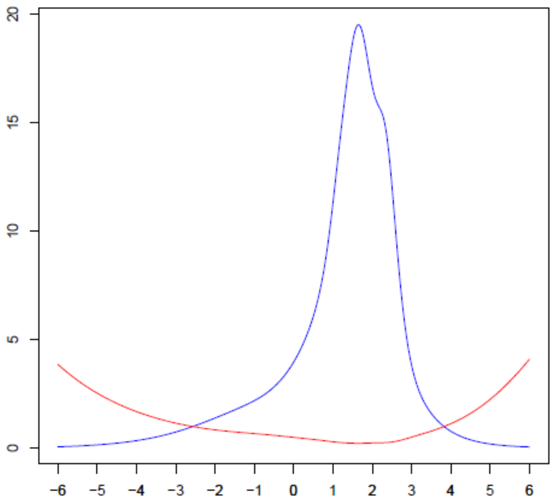

Fig. 6. TIF and SEM graphs

4. Conclusion

This study aims to evaluate the psychometric characteristics of reading ability items on the ProTEFL instrument using an IRT approach. To achieve this objective, a sample data feasibility check was carried out, IRT assumption testing, IRT model selection, and item parameter estimation were carried out. The results of the study indicate that the IRT 4PL model is the most suitable framework for analyzing ProTEFL reading comprehension items, as it provides richer diagnostic information than simpler models based on the AIC, BIC, log-likelihood, and standardized chi-square ($S\chi^2$) indicators. Of the total 50 items analyzed, 43 items were declared to fit the 4PL model and have met the assumptions of unidimensionality, local independence, and parameter invariance. The item parameters analyzed showed that the majority of items (70%) were within the ideal discrimination range, no items were too easy because the distribution of difficulty levels was quite spread (concentrated in the range of 0 to 2), most items had good pseudo-guessing parameter values because they were very low, and as many as 33% of items were indicated as not being able to be answered correctly by high-ability participants due to carelessness, question ambiguity, and other non-cognitive factors. It is recommended that ProTEFL test developers regularly conduct IRT-based item analyses to monitor item quality. Items with extreme discrimination indices, excessive difficulty, or low upper asymptote should be revised or replaced to maintain the fairness and accuracy of the test.

These findings have important implications for constructing competency-based English language proficiency test instruments and can be used to develop a standardized ProTEFL question bank. The IRT approach with the 4PL model has also been shown to capture phenomena such as identifying items that are likely to be incorrectly responded to by even highly proficient participants. However, this study has several limitations, such as the English language proficiency domain, which only covers reading and not the entire ProTEFL test instrument. Future studies should extend the analysis to other ProTEFL subdomains (listening and structure), to provide a more holistic psychometric evaluation. Further exploration of multidimensional IRT models or adaptive testing approaches may also offer deeper insights into language proficiency measurement.

Supplementary Materials: The following supporting table is available in the supplementary file; Table S1: *Item Quality Assessment Table Using the 4-Parameter Logistic IRT Model.*

Author Contribution: LA: Conceptualization, Writing – Original Draft, Formal Analysis, and Editing; MLBG: Methodology, Substantive Review, and Content Editing; ANR: Writing – Introduction Section, Language Translation, and Linguistic Editing. All authors have read and approved the final paper.

Funding: This research received no external funding.

Acknowledgment: The authors would like to thank Universitas Negeri Yogyakarta for the support and facilities provided throughout the research process. In particular, we appreciate the access granted to conduct comprehensive data analysis.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] M. Camilleri, "Higher education marketing communications in the digital era," in *Strategic Marketing of Higher Education in Africa*, pp. 77–95, 2020, https://doi.org/10.4324/9780429320934-7.

[2] N. D.-N. Nguyen, A. H.-C. Leung, H.-T. Hien, and N.-T. Thiet, "The impact of virtual exchanges using english as a lingua franca on students' linguistic and intercultural competence," *Teaching English as a Second or Foreign Language--TESL-EJ*, vol. 28, no. 3, 2024, https://doi.org/10.55593/ej.28111a1.

[3] S. Ehrenreich, "English as a business lingua franca in a German multinational corporation: Meeting the challenge," *Journal of Business Communication*, vol. 47, no. 4, pp. 408–431, 2010, https://doi.org/10.1177/0021943610377303.

[4] [4] J. Schmidgall, M. E. Oliveri, T. Duke, and E. Carter Grissom, "Justifying the construct definition for a new language proficiency assessment: the redesigned TOEIC Bridge® Tests - Framework paper," *ETS Research Report Series*, vol. 2019, no. 1, pp. 1–20, Dec. 2019, https://doi.org/10.1002/ets2.12267.

[5] [5] M. T. Hidayat, "English language proficiency and career opportunities: Perceptions of Indonesian university graduates," *Language Value*, vol. 17, no. 1, pp. 85–107, 2024, https://doi.org/10.6035/languagev.7933.

[6] S. Zein, D. Sukyadi, F. A. Hamied, and N. S. Lengkanawati, "English language education in Indonesia: A review of research (2011–2019)," *Language Teaching*, vol. 53, no. 4, pp. 491–523, 2020, https://doi.org/10.1017/S0261444820000208.

[7] R. A. Mustopa, V. S. Damamaianti, Y. Mulyati, and D. S. Anshori, "Challenges of reading literacy assessment in the digital age," pp. 408–416, 2024, https://doi.org/10.2991/978-94-6463-376-4_54.

[8] B. R. Rush, D. C. Rankin, and B. J. White, "The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value," *BMC Med Educ*, vol. 16, no. 1, p. 250, 2016, https://doi.org/10.1186/s12909-016-0773-3.

[9] M. L. B. Garcia, K. C. P. Santos, and C. P. Vistro-Yu, "Comparing two psychometric approaches: The case of item analysis for a classroom test in mathematics," *International Journal of Education and Practice*, vol. 13, no. 1, pp. 327–344, 2025, https://doi.org/10.18488/61.v13i1.4060.

[10] V. Mutiawani, A. M. Athaya, K. Saputra, and M. Subianto, "Implementing item response theory (IRT) method in quiz assessment system," *TEM Journal*, vol. 11, no. 1, pp. 210–218, 2022, https://doi.org/10.18421/TEM111-26.

[11] A. Gyamfi and R. Acquaye, "Parameters and models of item response theory (IRT): A review of literature," *Acta Educationis Generalis*, vol. 13, no. 3, pp. 68–78, 2023, https://doi.org/10.2478/atd-2023-0022.

[12] K. Hori, H. Fukuhara, and T. Yamada, "Item response theory and its applications in educational measurement Part I: Item response theory and its implementation in R," *WIREs Computational Statistics*, vol. 14, no. 2, pp. 1–22, 2022, https://doi.org/10.1002/wics.1531.

[13] A. Robitzsch, "On the choice of the item response model for scaling PISA data: Model selection based on information criteria and quantifying model uncertainty," *Entropy*, vol. 24, no. 6, p. 760, 2022, https://doi.org/10.3390/e24060760.

[14] Y.-L. Liaw, P. Baghaei, R. Strietholt, S. Meinck, and A. Strello, "Environmental knowledge: Conceptualization and measurement," in *Knowledge and Willingness to Act Pro-Environmentally*, vol. 16, pp. 31–56, 2025, https://doi.org/10.1007/978-3-031-76033-4_4.

[15] M. Ul Hassan and F. Miller, "Discrimination with unidimensional and multidimensional item response theory models for educational data," *Commun Stat Simul Comput*, vol. 51, no. 6, pp. 2992–3012, 2019, https://doi.org/10.1080/03610918.2019.1705344.

[16] R. Imtikhanah, E. Istiyono, and Widihastuti, "English final examination items befitting the criteria: An item response theory approach," *Eurasian Journal of Educational Research*, vol. 106, no. 106, pp. 338–350, 2023, https://openurl.ebsco.com/EPDB%3Agcd%3A16%3A14621552/detailv2?sid=ebsco%3Aplink%3Ascholar&id=ebsco%3Agcd%3A175313292&crl=c&link_origin=scholar.google.com.

[17] F. Antoniou, G. Alkhadim, A. Mouzaki, and P. Simos, "A psychometric analysis of raven's colored progressive matrices: evaluating guessing and carelessness using the 4pl item response theory model," *J Intell*, vol. 10, no. 1, p. 6, 2022, https://doi.org/10.3390/jintelligence10010006.

[18] Z. S. Ibrahim, H. Retnawati, A. Irambona, and B. E. O. Pérez, "Stability of estimation item parameter in IRT dichotomy considering the number of participants," *REID (Research and Evaluation in Education)*, vol. 10, no. 1, pp. 114–127, 2024, https://doi.org/10.21831/reid.v10i1.73055.

[19] L. Lina, D. Mardapi, and H. Haryanto, "Item characteristics on Pro-TEFL listening section," in *Proceedings of the First International Conference on Advances in Education, Humanities, and Language, ICEL 2019, Malang, Indonesia, 23-24 March 2019*, 2019, https://doi.org/10.4108/eai.11-7-2019.159630.

[20] Mohammad Umar Fakhrudin, P. W. Nugraha, N. Putri, and V. Nurfitriyani, "Investigating the bilingual education program at malhikdua school," *Journal of Literature Language and Academic Studies*, vol. 3, no. 3, pp. 105–109, 2024, https://doi.org/10.56855/jllans.v3i3.1288.

[21] A. Puente, A. P. Gutiérrez-de-Blume, J. Calderon, and L. Rojas, "Validación psicométrica de un test de comprensión lectora y conciencia metacognitiva en estudiantes universitarios," *Ocnos*, vol. 24, no. 2, 2025, https://doi.org/10.18239/ocnos_2025.24.2.538.

[22] T. Clark, L. Foster, and A. Bryman, "Analysing quantitative data," in *How to do your social research project or dissertation*, pp. 233–268, 2019, https://doi.org/10.1093/hepl/9780198811060.003.0013.

[23] H. Ahmad, N. Mamat, M. Che Mustafa, and S. Iryani Mohd Yusoff, "Validating the teaching, learning, and assessment quality of Malaysian ECCE instrument," *International Journal of Evaluation and Research in Education (IJERE)*, vol. 10, no. 1, p. 135, 2021, https://doi.org/10.11591/ijere.v10i1.20857.

[24] W. Guo and Y.-J. Choi, "Assessing dimensionality of IRT models using traditional and revised parallel analyses," *Educ Psychol Meas*, vol. 83, no. 3, pp. 609–629, 2023, https://doi.org/10.1177/00131644221111838.

[25] M. Gökcan and D. Çobanoğlu Aktan, "Validation of the vocabulary size test," *Egit Psikol Olcme Deger Derg*, vol. 13, no. 4, pp. 305–327, 2022, https://doi.org/10.21031/epod.1144808.

[26] R. K. Hambleton, H. Swaminathan, and H. Jane. Rogers, *Fundamentals of item response theory* (vol. 2). Sage Publications. 1991. https://books.google.co.id/books?hl=id&lr=&id=gW05DQAAQBAJ.

[27] D. Saepuzaman, H. Haryanto, Edi Istiyono, H. Retnawati, and Y. Yustiandi, "Analysis of items parameters on work and energy subtest using item response theory," *Jurnal Pendidikan MIPA*, vol. 22, no. 1, pp. 1–9, 2021, https://doi.org/10.23960/jpmipa/v22i1.pp1-9.

[28] M. Wilson, *Constructing measures: An item response modeling approach*. Routledge. 2023. https://doi.org/10.4324/9781003286929.

[29] D. O. Tobih, M. A. Ayanwale, O. A. Ajayi, and M. V. Bolaji, "The use of measurement frameworks to explore the qualities of test items," *International Journal of Evaluation and Research in Education (IJERE)*, vol. 12, no. 2, pp. 914–923, 2023, https://doi.org/10.11591/ijere.v12i2.23747.

[30] A. Buse, "The likelihood ratio, wald, and lagrange multiplier tests: An expository note," *Am Stat*, vol. 36, no. 3, pp. 153–157, 1982, https://doi.org/10.2307/2683166.

[31] M. Orlando and D. Thissen, "Likelihood-based item-fit Indices for dichotomous item response theory models," *Appl Psychol Meas*, vol. 24, no. 1, pp. 50–64, 2000, https://doi.org/10.1177/01466216000241003.

[32] M. A. Barton and F. M. Lord, "An upper asymptote for the three-parameter logistic item-response model," *ETS Research Report Series*, vol. 1981, no. 1, pp. i–8, 1981, https://doi.org/10.1002/j.2333-8504.1981.tb01255.x.

[33] C. DeMars, *Item response theory*. Oxford University Press, 2010. https://doi.org/10.1093/acprof:oso/9780195377033.001.0001.

[34] R. K. Hambleton and H. Swaminathan, *Item response theory: Principles and applications*. Springer Netherlands, 1985. https://doi.org/10.1007/978-94-017-1988-9.

[35] K. Świst, "Item analysis and evaluation using a four-parameter logistic model," *Edukacja*, vol. 3, no. 134, pp. 77–97, 2015, https://www.ceeol.com/search/article-detail?id=471062.

[36] A. Darmana, A. Sutiani, and Jasmidi, "Development of the thermochemistry- HOTS-tawheed multiple choice instrument," *J Phys Conf Ser*, vol. 1462, no. 1, p. 012057, 2020, https://doi.org/10.1088/1742-6596/1462/1/012057.

[37] N. Shrestha, "Factor analysis as a tool for survey Analysis," *Am J Appl Math Stat*, vol. 9, no. 1, pp. 4–11, 2021, https://doi.org/10.12691/ajams-9-1-2.

[38] R. Kumolohadi, F. Mangunsong, and J. Suleeman, "Development and validation of ethical awareness scale among university students," *Jurnal Pengukuran Psikologi dan Pendidikan Indonesia (JP3I)*, vol. 10, no. 1, pp. 20–31, 2021, https://doi.org/10.15408/jp3i.v10i1.17216.

[39] U. Lorenzo-Seva and P. J. Ferrando, "MSA: The forgotten index for identifying inappropriate items before computing exploratory item factor analysis," *Methodology*, vol. 17, no. 4, pp. 296–306, 2021, https://doi.org/10.5964/meth.7185.

[40] N. D. Astuti, M. Hajaroh, Y. Prihatni, A. Setiawan, F. A. Setiawati, and H. Retnawati, "Comparison of KMO results, eigen value, reliability, and standard error of measurement: Original & rescaling through summated rating scaling," *JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia)*, no. 2, pp. 199–215, 2024, https://doi.org/10.15408/jp3i.v13i2.36684.

[41] P. H. Santoso, E. Istiyono, and H. Haryanto, "Principal component analysis and exploratory factor analysis of the mechanical waves conceptual survey," *JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia)*, vol. 11, no. 2, pp. 209–225, 2022, https://doi.org/10.15408/jp3i.v11i2.27488.

[42] L. Arriza, H. Retnawati, and R. T. Ayuni, "Item analysis of high school specialization mathematics exam questions with item response theory approach," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 18, no. 1, pp. 0151–0162, Mar. 2024, https://doi.org/10.30598/barekengvol18iss1pp0151-0162.

[43] A. Akbari, "The rasch analysis of item response theory: An untouched area in evaluating student academic translations," *SKASE Journal of Translation and Interpretation*, vol. 18, no. 1, pp. 50–77, 2025, https://doi.org/10.33542/JTI2025-1-05.

[44] W. Astuti and Adiwijaya, "Support vector machine and principal component analysis for microarray data classification," *J Phys Conf Ser*, vol. 971, p. 012003, 2018, https://doi.org/10.1088/1742-6596/971/1/012003.

[45] S. J. Haberman, S. Sinharay, and K. H. Chon, "Assessing item fit for unidimensional item response theory models using residuals from estimated item response functions," *Psychometrika*, vol. 78, no. 3, pp. 417–440, 2013, https://doi.org/10.1007/s11336-012-9305-1.

[46] E. Moradi, Z. Ghabanchi, and R. Pishghadam, "Reading comprehension test fairness across gender and mode of learning: Insights from IRT-based differential item functioning analysis," *Language Testing in Asia*, vol. 12, no. 1, p. 39, 2022, https://doi.org/10.1186/s40468-022-00192-3.

[47] S. Min and V. Aryadoust, "A systematic review of item response theory in language assessment: Implications for the dimensionality of language ability," *Studies in Educational Evaluation*, vol. 68, p. 100963, Mar. 2021, https://doi.org/10.1016/j.stueduc.2020.100963.

[48] Y. Liu and A. Maydeu-Olivares, "Local dependence diagnostics in IRT modeling of binary data," *Educ Psychol Meas*, vol. 73, no. 2, pp. 254–274, 2013, https://doi.org/10.1177/0013164412453841.

[49] S. Noventa, S. Ye, A. Kelava, and A. Spoto, "On the identifiability of 3- and 4-parameter item response theory models from the perspective of knowledge space theory," *Psychometrika*, vol. 89, no. 2, pp. 486–516, 2024, https://doi.org/10.1007/s11336-024-09950-z.

[50] G. Rasch, *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research, 1993, https://eric.ed.gov/?id=ED419814.

[51] A. Birnbaum, "Some latent trait models and their use in inferring an examinee's ability," in *Statistical theories of mental test scores*, ch. 5, p. 396, 1968, https://cir.nii.ac.jp/crid/1571698600108572800.

[52] M. A. Sorrel, F. J. Abad, J. Olea, J. de la Torre, and J. R. Barrada, “Inferential item-fit evaluation in cognitive diagnosis modeling,” *Appl Psychol Meas*, vol. 41, no. 8, pp. 614–631, 2017, https://doi.org/10.1177/0146621617707510.

[53] Z. Han, S. Sinharay, M. S. Johnson, and X. Liu, “The standardized S-X2 statistic for assessing item fit,” *Appl Psychol Meas*, vol. 47, no. 1, pp. 3–18, 2023, https://doi.org/10.1177/01466216221108077.

[54] X. Meng, G. Xu, J. Zhang, and J. Tao, “Marginalized maximum a posteriori estimation for the four-parameter logistic model under a mixture modelling framework,” *British Journal of Mathematical and Statistical Psychology*, vol. 73, no. S1, pp. 51–82, 2020, https://doi.org/10.1111/bmsp.12185.

[55] N. G. Waller and L. Feuerstahler, “Bayesian modal estimation of the four-parameter item response model in real, realistic, and idealized data sets,” *Multivariate Behav Res*, vol. 52, no. 3, pp. 350–370, 2017, https://doi.org/10.1080/00273171.2017.1292893.

[56] D. Magis, “A note on the item information function of the four-parameter logistic model,” *Appl Psychol Meas*, vol. 37, no. 4, pp. 304–315, 2013, https://doi.org/10.1177/0146621613475471.

[57] J. Jumini and H. Retnawati, “Estimating item parameters and student abilities: An IRT 2PL analysis of mathematics examination,” *AL-ISHLAH: Jurnal Pendidikan*, vol. 14, no. 1, pp. 385–398, 2022, https://doi.org/10.35445/alishlah.v14i1.926.

[58] H. Chin, C. M. Chew, W. Yew, and M. Musa, “Validating the cognitive diagnostic assessment and assessing students’ mastery of ‘parallel and perpendicular lines’ using the rasch model,” *Participatory Educational Research*, vol. 9, no. 6, pp. 436–452, 2022, https://doi.org/10.17275/per.22.147.9.6.

[59] Herwin and S. C. Dahalan, “Person fit statistics to identify irrational response patterns for multiple-choice tests in learning evaluation,” *Pegem Journal of Education and Instruction*, vol. 12, no. 4, pp. 39–46, 2022, https://doi.org/10.47750/pegegog.12.04.05.

[60] E. Ulitzsch, S. N. Yildirim-Erbasli, G. Gorgun, and O. Bulut, “An explanatory mixture IRT model for careless and insufficient effort responding in self-report measures,” *British Journal of Mathematical and Statistical Psychology*, vol. 75, no. 3, pp. 668–698, 2022, https://doi.org/10.1111/bmsp.12272.

[61] S. Gören, H. Kara, B. Erdem Kara, and H. Kelecioğlu, “The effect of aberrant responses on ability estimation in computer adaptive tests,” *Journal of Measurement and Evaluation in Education and Psychology*, vol. 13, no. 3, pp. 256–268, 2022, https://doi.org/10.21031/epod.1067307.

[62] W.-W. Liao, R.-G. Ho, Y.-C. Yen, and H.-C. Cheng, “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*, vol. 40, no. 10, pp. 1679–1694, 2012, https://doi.org/10.2224/sbp.2012.40.10.1679.

[63] A. L. Nichols and J. E. Edlund, “Why don’t we care more about carelessness? Understanding the causes and consequences of careless participants,” *Int J Soc Res Methodol*, vol. 23, no. 6, pp. 625–638, 2020, https://doi.org/10.1080/13645579.2020.1719618.

[64] E. Loken and K. L. Rulison, “Estimation of a four-parameter item response theory model,” *British Journal of Mathematical and Statistical Psychology*, vol. 63, no. 3, pp. 509–525, 2010, https://doi.org/10.1348/000711009X474502.