VELES JOURNAL

Vol. 9, No. 1; April 2025
E-ISSN 2579-7484
Pages 218-230
DOI: http://dx.doi.org/10.29408/veles.v9i1.27644

Evaluating English Language Test Items Developed by Teachers: An Item Response Theory Approach

***[1]Rezkilaturahmi, [2]Muhammad Istiqlal, [1]Nur Hidayanto Pancoro Setyo Putro, [1]Edi Istiyono, [1]Widihastuti**
[1]Universitas Negeri Yogyakarta, Indonesia
[2]UIN Salatiga, Indonesia

***Correspondence:**
rezkilaturahmi.2023@student.uny.ac.id

Submission History:
Submitted: October 1, 2024
Revised: April 1, 2025
Accepted: April 22, 2025

Abstract

Evaluating students' abilities in educational settings is crucial for assessing learning outcomes and instructional effectiveness. In Indonesia, many schools have developed local English language assessments, yet these tests often lack psychometric validation. This study aims to evaluate the quality of a teacher-developed English language test instrument using the Item Response Theory (IRT) approach. A total of 25 multiple-choice items created by the English teacher group in Muna Regency were administered to 162 students from five randomly selected schools. A descriptive quantitative method was employed with the aid of R Studio for data analysis. Initial sample adequacy was confirmed using the Kaiser-Meyer-Olkin (KMO = 0.686) and Bartlett's Test of Sphericity ($p < .001$). The study applied model fit analyses for 1-PL, 2-PL, and 3-PL logistic models, with the 2-PL model emerging as the most appropriate, as 16 items demonstrated good fit. Further analysis of item characteristics under the 2-PL model revealed that only 11 items had acceptable difficulty and discrimination indices. In comparison, the remaining 14 items were either too easy, too complex, or poorly discriminating. These results indicate that a substantial portion of the test requires revision. The study highlights the importance of psychometric evaluation in teacher-made assessments and recommends capacity-building for teachers in test development and validation practices.

Keywords: Item response theory, English test, item difficulty, English language assessments.

INTRODUCTION

Assessment is a fundamental component of the teaching and learning process, serving as a systematic means to collect, interpret, and utilise data on student performance (Wiliam, 2011; Earl, 2013). In the context of English language education, Durán (2008) argue that assessments not only provide evidence of learners' language proficiency but also offer valuable insights into the effectiveness of instructional practices. Through well-designed assessments, educators can determine whether students have achieved the intended learning outcomes and identify specific areas of strength and weakness (Banta & Palomba,

2015; Ali, 2018; McTighe & Ferrara, 2021). Moreover, assessment data play a pivotal role in informing broader educational decisions. Analysing the patterns in student performance, teachers and school administrators can make informed judgments about the appropriateness of teaching materials (Sharma, 2015), the pace of instruction (Sun et al., 2016), and the overall alignment between curriculum goals and student achievement (DeLuca & Bellara, 2013). In this way, assessment becomes a diagnostic tool that not only evaluates student learning but also drives pedagogical improvement.

In many educational contexts, teachers are frequently required to design their assessment instruments (Akram & Zepeda, 2015; Borg & Edmett, 2018; Kasman & Lubis, 2022). This practice often stems from practical demands, such as aligning tests with locally implemented curricula (English, 2010), addressing specific classroom learning objectives, or compensating for the absence of standardised test materials (Vatterott, 2015). Teacher-developed tests can offer several advantages, including contextual relevance, flexibility in content delivery, and immediate responsiveness to student needs. However, despite their practicality, many teacher-made tests suffer from a lack of empirical validation (Hartell & Strimel, 2018; Hirpassa, 2018; Gashaye & Degwale, 2019; Effendi & Mayuni, 2022). Unlike standardised assessments, which undergo rigorous procedures to establish reliability, validity, and fairness, classroom-based tests are often constructed without the benefit of psychometric analysis or peer review (Areekkuzhiyil, 2021). The absence of validation raises critical concerns about the accuracy and fairness of the test results, particularly when these scores are used to make high-stakes decisions about student achievement, placement, or promotion (Shaw et al., 2012). Furthermore, the reliance on intuition and experience rather than data-driven evaluation can undermine the credibility of teacher-made tests (Young & Kim, 2010; Sundqvist et al., 2017).

Given these challenges, there is a pressing need for more rigorous methods to evaluate the quality of teacher-made assessments. Item Response Theory (IRT) is a modern psychometric approach that provides a sophisticated framework for evaluating individual test items based on their statistical performance (Zanon et al., 2016; Reeve, 2023; Wilson, 2023). Unlike Classical Test Theory (CTT), which treats test scores as the sum of observed responses and attributes measurement error uniformly across all items, IRT focuses on the interaction between item characteristics and a test-taker's underlying ability. Two key parameters estimated in IRT are item difficulty and item discrimination (Sweeney et al., 2022). The difficulty parameter indicates the level of ability a student must possess to have a 50% chance of answering the item correctly, while the discrimination parameter reflects how well an item differentiates between high- and low-performing students (Wauters et al., 2010; Lee, 2019). One of the primary advantages of IRT over CTT is its sample-independent nature; the parameters estimated under IRT are considered stable across different populations, assuming model fit is achieved (Embretson & Yang, 2006). Furthermore, IRT enables the identification of poorly functioning items, such as those that are too easy, too difficult, or fail to discriminate effectively, which can then be revised or removed to improve the overall validity and reliability of the test (Lord, 1980; Haladyna & Rodriguez, 2013).

Numerous studies have emphasised the critical role of item analysis in evaluating the quality of teacher-made English tests, particularly with regard to item difficulty, discrimination, validity, and reliability. Karim et al. (2021) revealed that many English teachers do not perform item analysis, leading to the uncritical use of test items with poor

difficulty and discrimination indices. Their study found that only a small proportion of items in a reading comprehension test met acceptable standards, while the majority required revision or rejection. Similarly, Darmawan et al. (2022) investigated a high school English test. They found a wide variation in item difficulty and discrimination levels, noting that while some items were of good quality, many were invalid or insufficiently discriminating, thereby warranting further evaluation. Complementing these findings, Effendi and Mayuni (2022) argued that although teacher-made multiple-choice tests are widely used due to their practical relevance, continuous quality monitoring is essential. Their item response analysis indicated that the test could have been more credible with appropriate revisions, leading them to recommend institutional training in test development and validation.

Maharani and Putro (2020) also highlighted the importance of item analysis in their study of final semester tests in East Java. While their analysis revealed a high proportion of items with excellent discrimination power and effective distractors, the test lacked a balanced distribution across difficulty levels, underscoring the need for more systematic test construction. In a broader evaluative context, Wuntu (2021) analysed summative tests over two academic semesters and found both tests to be valid, highly reliable, and generally effective in item discrimination and facility, with most items recommended for continued use. Collectively, these studies affirm the necessity of conducting empirical item analysis to ensure the fairness, accuracy, and instructional value of teacher-made assessments in English language education.

While classroom assessments are central to measuring students' learning outcomes and informing instructional decisions, the quality of teacher-developed tests remains an area of concern, particularly in regional educational contexts where standardised instruments are scarce. Despite growing recognition of the importance of empirical validation, many locally constructed English tests continue to be administered without rigorous item-level analysis, limiting their diagnostic and evaluative utility. Existing research has primarily focused on classical test approaches or has been concentrated in urban or institutionally resourced areas, leaving the psychometric evaluation of rural or underrepresented regions relatively underexplored. Moreover, the application of modern measurement models such as Item Response Theory (IRT) in the context of teacher-made English tests remains limited, especially when it comes to systematically assessing item difficulty and discrimination parameters across diverse learner populations. By applying IRT to a test developed collaboratively by English teachers in multiple schools in Muna Regency, Indonesia, this study provides a more nuanced understanding of item performance. It contributes to enhancing assessment literacy among practitioners in remote educational settings. The findings are expected not only to identify strengths and weaknesses of the test items but also to support the development of more valid and equitable assessments in similar local contexts.

METHOD

This study employed a quantitative research design, which is appropriate for systematically measuring and analysing numerical data related to test item quality (Creswell, 2014). Specifically, the study focused on conducting item-level analysis of a teacher-developed English language test using Item Response Theory (IRT). The quantitative approach enables objective evaluation of test item characteristics such as

difficulty and discrimination, which are essential for establishing test validity and fairness (Ary et al., 2019). The design was structured to support empirical analysis of item performance using statistical tools within the IRT framework.

A total of 25 multiple-choice cognitive items, designed by a group of English teachers in Muna Regency, Southeast Sulawesi, Indonesia, were administered to 162 students from five randomly selected schools. The items were intended to assess students' English language proficiency, with a particular focus on reading comprehension. Before the test was administered, the items underwent a rigorous expert-based content validation process, involving five specialists in English language education. Each item was reviewed across three key dimensions: content relevance, to ensure alignment with the learning objectives; item construction, to verify adherence to sound item-writing principles; and language clarity, to confirm that the language used was accessible and unambiguous for students. This validation process followed well-established item development frameworks (Nitko & Brookhart, 2014). To further establish the instrument's validity, Aiken's V index was employed, yielding a coefficient of 0.96, which reflects a high level of expert agreement and confirms the instrument's strong content validity (Aiken, 1985).

Following data collection, the responses were processed and analysed using R Studio software, a statistical computing environment suitable for psychometric evaluation. Prior to IRT modelling, a model fit analysis was conducted to determine whether the data met the assumptions required for applying a unidimensional IRT model. The analysis proceeded with the application of the 2-Parameter Logistic Model (2-PL), which evaluates each item based on difficulty (b parameter) and discrimination (a parameter). The instrument's internal consistency reliability was also estimated using R, yielding a reliability coefficient of 0.787, suggesting acceptable consistency for classroom-level assessments (George & Mallery, 2003). This study provided a detailed evaluation of how well individual test items functioned in distinguishing between students of varying proficiency levels, offering actionable insights for improving test quality.

FINDING AND DISCUSSION

The suitability of the Item Response Theory (IRT) model was evaluated by examining the chi-square values for each item under the 1-Parameter Logistic (1pl), 2-Parameter Logistic (2pl), and 3-Parameter Logistic (3pl) models. The model fit analysis was conducted using R Studio, where the obtained chi-square statistics for each item were compared against the corresponding critical chi-square values from the distribution table. This comparison determined whether each item appropriately fit the assumptions of the respective IRT models. The detailed results of the model fit analysis for all 25 items are presented in the following table.

Table 1. Sample adequacy in KMO and Barlett's test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.686
Bartlett's Test of Sphericity	Approx. Chi-Square	269.332
	Df	45
	Sig	.000

The results of the sample adequacy test are presented in Table 1. The Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy produced a value of 0.686, which indicates a

moderate level of sampling adequacy. According to Kaiser (1974) as cited in Nkansah (2018), a KMO value between 0.60 and 0.70 is considered acceptable for proceeding with further item-level analysis. This suggests that the sample used in this study was sufficient for conducting Item Response Theory (IRT) analysis. In addition, Bartlett's Test of Sphericity yielded a chi-square value of 269.332 with 45 degrees of freedom, and a significance level of $p = 0.000$. The significant result ($p < 0.05$) confirms that there are adequate correlations among the items, meaning that the data are suitable for factor-based or item-level modelling. These findings provide statistical evidence that the data meet the assumptions required for applying IRT, thereby justifying the use of this model for evaluating the quality of the teacher-made test items.

Fit Model of the English Test

The model fit analysis using logistic parameter models—namely the 1-Parameter Logistic (1pl), 2-Parameter Logistic (2pl), and 3-Parameter Logistic (3pl) models—was conducted to determine the appropriateness of each item under Item Response Theory (IRT). The chi-square (χ^2) goodness-of-fit values for all 25 items were compared with the corresponding critical chi-square values based on their degrees of freedom. This analysis was performed using R Studio, and the results are presented in Table 2.

Table 2. Result of logistic parameter model fit analysis

NS	1-PL				2-PL				3-PL			
	x^2	dk	x^2_{kritis}	Cat	x^2	dk	x^2_{kritis}	Cat	x^2	dk	x^2_{kritis}	Cat
1.	37,3	9	16,9	-	18,4	12	21,02	Fit	16,9	10	18,3	Fit
2.	20,1	7	14,01	-	18,5	8	15,5	-	18,09	6	12,5	-
3.	26,7	10	18,3	-	9,88	13	22,3	Fit	8,51	12	21,02	Fit
4.	15,3	11	19,6	Fit	21,3	13	22,3	Fit	18,7	12	21,02	Fit
5.	13,2	12	21,02	Fit	17,8	11	19,6	Fit	18,3	10	18,3	-
6.	13,8	11	19,6	Fit	4,91	10	18,3	Fit	4,56	7	14,01	Fit
7.	16,4	12	21,02	Fit	20,9	12	21,02	Fit	17,7	11	19,6	Fit
8.	26,7	12	21,02	-	21,7	3	7,81	-	26,03	3	7,81	-
9.	10,8	10	18,3	Fit	13,4	10	18,3	Fit	11,1	9	16,9	Fit
10.	21,9	11	19,6	-	57,6	5	11,07	-	22,6	5	11,07	-
11.	11,9	11	19,6	Fit	20,2	10	18,3	-	38,6	8	15,5	-
12.	14,7	10	18,3	Fit	20,7	12	21,02	Fit	17,9	10	18,3	Fit
13.	17,7	9	16,9	-	10,7	13	22,3	Fit	9,30	10	18,3	Fit
14.	19,4	9	16,9	-	9,93	10	18,3	Fit	9,71	9	16,9	Fit
15.	6,15	12	21,02	Fit	5,05	12	21,02	Fit	6,51	11	19,6	Fit
16.	13,8	11	19,6	Fit	4,91	10	18,3	Fit	4,56	7	14,01	Fit
17.	16,4	12	21,02	Fit	20,9	12	21,02	Fit	17,7	11	19,6	Fit
18.	26,7	12	21,02	-	21,7	3	7,81	-	26,03	3	7,81	-
19.	11,02	10	18,3	Fit	15,08	10	18,3	Fit	11,8	9	16,9	Fit
20.	40,2	11	19,6	-	30,0	13	22,3	-	27,5	11	19,6	-
21.	11,9	11	19,6	Fit	20,2	10	18,3	-	38,6	8	15,5	-
22.	14,7	10	18,3	Fit	20,7	12	21,02	Fit	17,9	10	18,3	Fit
23.	17,7	9	16,9	-	10,7	13	22,3	Fit	9,30	10	18,3	Fit
24.	40,7	10	16,9	-	23,4	13	22,3	-	23,7	11	19,6	-
25.	21,9	11	19,6	-	57,6	5	11,07	-	22,6	5	11,07	-
Count		1PL		13		2PL		16		3PL		15

The findings reveal that the 2pl model yielded the highest number of items with acceptable fit, with 16 out of 25 items falling within the critical value threshold and therefore considered "Fit." The 3pl model followed closely with 15 items fitting the model, while the 1pl model showed only 13 items fitting the assumptions. These results suggest that the 2-Parameter Logistic model is the most appropriate for analysing the test, as it balances both item difficulty and item discrimination more effectively than the simpler 1pl model or the more complex 3pl model. Moreover, several items did not meet the chi-square fit criteria in any of the models, indicating potential issues such as misfitting response patterns or problematic item characteristics. Items such as 1, 2, 8, 10, 20, 24, and 25 were consistently misfitting across all models, suggesting that these items may require revision or removal from the test to improve the overall instrument quality.

Table 3. 2-PL based on unidimensional dichotomous scoring

S	a Parameter	b Parameter	Category
1.	-0,412	-3.902	-
2.	-0,517	-4.448	-
3.	-0.242	-1.867	Good
4.	-0.032	-25.980	-
5.	1.135	-0.263	Good
6.	1.564	-0.666	Good
7.	0.136	-0.517	Good
8.	59.735	-0.502	-
9.	0.830	0.714	Good
10.	5.702	0.335	-
11.	1.168	0.414	Good
12.	0.046	25.191	-
13.	-0.293	-5.121	-
14.	-0.110	-15.855	-
15.	0.585	-1.786	Good
16.	1.564	-0.666	Good
17.	0.136	-0.517	Good
18.	59.735	-0.502	-
19.	0.925	0.590	Good
20.	0.190	4.163	-
21.	1.168	0.414	Good
22.	0.046	25.191	-
23.	-0.293	-5.121	-
24.	-0.028	-15.900	-
25	5.702	0.335	-

Table 3 shows the results of item parameter estimation using the 2-Parameter Logistic (2pl) model based on unidimensional dichotomous scoring. Each item was evaluated in terms of its discrimination index (a parameter) and difficulty level (b parameter). An item is considered acceptable if its discrimination index falls within the range of 0 to 2 and its difficulty level is within the range of -2 to +2.

Based on these criteria, 11 out of 25 items were categorised as “Good”, meaning they met both parameter thresholds. These items functioned effectively in distinguishing between students of varying proficiency levels and those who had appropriate levels of difficulty. The items that met these criteria were items 3, 5, 6, 7, 9, 11, 15, 16, 17, 19, and 21.

In contrast, 14 items did not meet the required standards. Several of these items had values that were either too extreme or implausible. For instance, items such as 8 and 18 showed very high discrimination values, which may indicate irregular response patterns or statistical anomalies. Other items, like 2, 4, 12, 14, and 24, displayed very high or very low difficulty levels, suggesting that they may not be suitable for accurately assessing students' English proficiency. These results indicate that while a portion of the teacher-made test items are functioning well, a significant number require revision. This highlights the importance of item analysis in improving the overall quality of test instruments and ensuring that assessments provide meaningful and accurate information about student performance.

Invariance of 2-PL

A fundamental strength of the 2-Parameter Logistic (2pl) model lies in its assumption of parameter invariance—the idea that item parameters, such as difficulty (b) and discrimination (a), should remain stable across different samples or subgroups. This property ensures that the quality of test items is not dependent on a particular group of test-takers, thereby enhancing the fairness and generalizability of the assessment instrument. To examine this property, an odd-even split analysis was conducted, dividing the dataset into two subgroups based on the order of respondents (odd-numbered and even-numbered). The 2pl model was then applied separately to each subset, and the resulting item parameter estimates were compared through scatter plots shown in Figure 1.

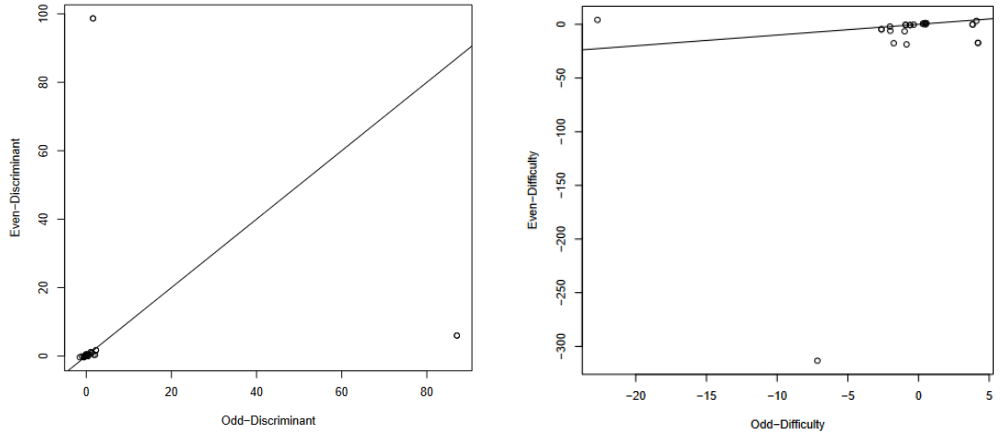

Figure 1. Invariance of 2-PL

Figure 1 displays two scatter plots. The top plot compares the discrimination indices (a-parameters) between the odd and even subgroups. In contrast, the bottom plot compares the difficulty levels (b-parameters) for the same items across the two groups. The diagonal line in each plot represents perfect agreement between the two estimates.

In the discrimination plot, most data points are clustered near the diagonal line, indicating that the majority of items yielded consistent discrimination values across both subgroups. This suggests that the ability of these items to differentiate between high- and low-performing students was relatively stable. However, two items appear as clear outliers, with unusually high discrimination values in one subgroup but not in the other, signalling a violation of invariance. These outlier items may reflect irregular student response patterns or problematic item construction. Besides, in the difficulty plot, a similar pattern emerges: most items are grouped near the expected range with consistent difficulty values between

subgroups. However, several items display extreme deviations, particularly with difficulty parameters well outside the recommended range (e.g., below –250), suggesting that these items functioned differently across samples. Such inconsistency indicates unstable item difficulty, which could compromise the reliability of the test.

DISCUSSION

This study investigated the quality of a teacher-developed English language test using the 2-Parameter Logistic (2pl) model of Item Response Theory (IRT), providing evidence-based insights into item performance and model suitability. The findings emphasise the importance of psychometric validation in classroom assessments, particularly those developed by teachers without formal measurement training.

Item Response Theory offers multiple logistic models depending on the number of item parameters included. The 1-PL model considers only item difficulty, the 2-PL model adds item discrimination, and the 3-PL model incorporates a pseudo-guessing parameter (Embretson & Reise, 2000; Baldonado et al., 2015; Bichi & Talib, 2018). In this study, model fit analysis using chi-square comparisons revealed that 13 items fit the 1-PL model, 16 items fit the 2-PL model, and 15 items fit the 3-PL model. Among these, the 2-PL model demonstrated the highest number of well-fitting items and was therefore selected for further analysis. This aligns with recommendations in educational measurement literature, where the 2-PL model is frequently preferred when both item difficulty and discrimination are relevant to instructional decision-making (Setiawati et al., 2018; Immekus et al., 2019).

The application of the 2-PL model allowed for an in-depth evaluation of each item based on its discrimination index (a parameter) and difficulty level (b parameter). According to established psychometric criteria, a well-functioning item typically has a discrimination value between 0.0 and 2.0 and a difficulty value within the range of -2.0 to +2.0 (De Ayala, 2009; Baker, 2001). The analysis revealed that 11 out of the 25 items met these thresholds and were categorised as "Good." These items were effective in distinguishing students with varying ability levels and fell within a difficulty range appropriate for the target population.

However, 14 items were identified as problematic, with some showing extreme values that exceeded acceptable thresholds, such as discrimination indices above 2.0 or difficulty values far outside the -2.0 to +2.0 range. These values suggest either item malfunction or poor alignment with the test-takers' proficiency levels. Items with excessively high discrimination may indicate over-sensitivity to minor ability differences, potentially leading to biased results (Osterlind, 2006). Conversely, items with extreme difficulty values may be too easy or too hard for the test population, reducing their diagnostic value and potentially introducing ceiling or floor effects (Brown, 2005; Odukoya et al., 2017; Metsämuuronen, 2022).

The invariance analysis, conducted through an odd-even split of the dataset, provided additional insights into the stability of parameter estimates. Scatter plots comparing item parameters across the two subgroups showed that most items exhibited consistent values for both difficulty and discrimination. This suggests general parameter invariance and supports the robustness of the 2pl model when applied to teacher-made assessments (Carlson & von Davier, 2017; Brown & Abdulnabi, 2017). However, several items showed substantial discrepancies across subgroups, appearing as outliers in the scatter plots. Such inconsistencies raise concerns about the fairness of those items and suggest potential bias

or sensitivity to subgroup characteristics (Zieky, 2016). These items should be revised, excluded, or further examined before future test use.

The findings underscore the importance of empirical item validation in teacher-made assessments. While multiple-choice formats remain popular due to their practicality and ease of scoring, they are prone to psychometric flaws when not properly analysed (Nitko & Brookhart, 2014). This study shows that integrating IRT-based techniques into assessment design—primarily through tools like R Studio—can substantially enhance the quality, fairness, and interpretive value of classroom tests. Moreover, the results affirm that valid and reliable test development is achievable in decentralised educational settings like Muna Regency, provided that educators are supported with the right tools and training. In line with this, professional development in test construction and item analysis should be prioritised for in-service and pre-service teachers (Kissi et al., 2023). Institutions and policymakers must promote assessment literacy to bridge the gap between curriculum delivery and test quality, ensuring that classroom-based evaluations reflect accurate and equitable measures of student learning outcomes (Care et al., 2012; Pastore, 2023).

CONCLUSION

This study aimed to evaluate the quality of a teacher-developed English language test by applying the 2-Parameter Logistic (2pl) model of Item Response Theory (IRT). The research focused on analysing item characteristics—particularly item difficulty and discrimination—and assessing the model's suitability and parameter stability across subgroups of students. The findings revealed that the 2pl model provided the best fit for the test data compared to the 1-PL and 3-PL models, with 16 out of 25 items showing statistical conformity under the 2pl framework. Further analysis indicated that 11 items met the established psychometric criteria for both difficulty and discrimination, suggesting that these items functioned well in differentiating between students of varying ability levels. Conversely, 14 items exhibited weaknesses, such as extreme parameter values or instability across subsamples, and were thus classified as requiring revision or removal.

An odd-even split analysis confirmed partial parameter invariance, with most items demonstrating consistent estimates across groups. However, a few items appeared as outliers, highlighting the need for careful item review to ensure fairness and generalizability. These results affirm the utility of the 2pl model in identifying both strengths and weaknesses in teacher-made assessments and underscore the importance of data-driven validation in classroom testing practices. In light of these findings, it is concluded that empirical item analysis is essential for ensuring test validity, reliability, and fairness, especially in educational settings where teachers independently develop their assessments.

REFERENCES

Aiken, L. R. (1985). Three coefficients for analysing the reliability and validity of ratings. *Educational and Psychological Measurement, 45*(1), 131–142. https://doi.org/10.1177/0013164485451012

Akram, M., & Zepeda, S. J. (2015). Development and validation of a teacher self-assessment instrument. *Journal of Research and Reflections in Education*, *9*(2).

Ali, L. (2018). The design of curriculum, assessment, and evaluation in higher education should be done with constructive alignment. *Journal of Education and e-Learning Research*, *5*(1), 72–78. https://doi.org/10.20448/journal.509.2018.51.72.78

Areekkuzhiyil, S. (2021). Issues and concerns in classroom assessment practices. *Edutracks*, *20*(8), 20-23.

Ary, D., Jacobs, L. C., Irvine, C. K. S., & Walker, D. (2019). *Introduction to research in education* (10th ed.). Cengage Learning.

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

Baldonado, A. A., Svetina, D., & Gorin, J. (2015). Using necessary information to identify item dependence in Passage-Based Reading Comprehension tests. *Applied Measurement in Education*, *28*(3), 202–218. https://doi.org/10.1080/08957347.2015.1042154

Banta, T. W., & Palomba, C. A. (2015). *Assessment essentials: Planning, implementing, and improving assessment in higher education* (2nd ed.). Jossey-Bass/Wiley.

Bichi, A. A., & Talib, R. (2018). Item Response Theory: An introduction to latent trait models for test and item development. *International Journal of Evaluation and Research in Education (IJERE)*, *7*(2), 142. https://doi.org/10.11591/ijere.v7i2.12900

Borg, S., & Edmett, A. (2018). Developing a self-assessment tool for English language teachers. *Language Teaching Research*, *23*(5), 655–679. https://doi.org/10.1177/1362168817752543

Brown, H. D. (2005). *Language assessment: Principles and classroom practices*. Pearson Education.

Brown, G. T. L., & Abdulnabi, H. H. A. (2017). Evaluating the quality of higher education instructor-constructed multiple-choice tests: Impact on student grades. *Frontiers in Education*, *2*. https://doi.org/10.3389/feduc.2017.00024

Care, E., Griffin, P., & McGaw, B. (2012). *Assessment and teaching of 21st century skills* (pp. 17-66). Dordrecht, The Netherlands: Springer.

Carlson, J.E., & von Davier, M. (2017). Item Response Theory. In: Bennett, R., von Davier, M. (eds) *Advancing human assessment. Methodology of educational measurement and assessment*. Springer, Cham. https://doi.org/10.1007/978-3-319-58689-2_5

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). SAGE Publications.

Darmawan, N. M., Sudarsono, Riyanti, N. D., Yuliana, N. Y. G. S., & Sumarni, N. (2022). Test items analysis of the English teacher-made test. *Journal of English Education and Teaching*, *6*(4), 498–513. https://doi.org/10.33369/jeet.6.4.498-513

de Ayala, R. J. (2009). *The theory and practice of item response theory.* Guilford Press.

DeLuca, C., & Bellara, A. (2013). The current state of assessment education. *Journal of Teacher Education*, *64*(4), 356–372. https://doi.org/10.1177/0022487113488144

Durán, R. P. (2008). Assessing English-language learners' achievement. *Review of Research in Education*, *32*(1), 292–327. https://doi.org/10.3102/0091732x07309372

Earl, L. M. (2013). *Assessment as learning: Using classroom assessment to maximise student learning* (2nd ed.). Corwin Press.

Effendi, T., & Mayuni, I. (2022). Examining a teacher-made English test in a language school. *LADU Journal of Languages and Education*, *2*(2), 67–76. https://doi.org/10.56724/ladu.v2i2.109

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Lawrence Erlbaum Associates Publishers.

English, F. W. (2010). *Deciding what to teach and test: Developing, aligning, and leading the curriculum* (3rd ed.). Corwin Press.

Embretson, S., & Yang, X. (2006). Item response theory. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 385–409). Mahwah, NJ: Lawrence Erlbaum Associates.

Gashaye, S., & Degwale, Y. (2019). The content validity of high school English language teacher-made tests: The case of Debre Work Preparatory School, East Gojjam, Ethiopia. *International Journal of Research in Engineering, IT and Social Sciences*, *9*(11), 41-50.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference* (4th ed.). Allyn & Bacon.

Gryphon, P., & Care, E. (2014). *Assessment and teaching of 21st century skills: Methods and Approaches*. Springer.

Haladyna, T.M., & Rodriguez, M.C. (2013). *Developing and validating test items* (1st ed.). Routledge. https://doi.org/10.4324/9780203850381

Hartell, E., & Strimel, G. J. (2018). What is it called and how does it work: Examining content validity and item design of teacher-made tests. *International Journal of Technology and Design Education*, *29*(4), 781–802. https://doi.org/10.1007/s10798-018-9463-2

Hirpassa, M. (2018). Content validity of EFL teacher-made assessment: The case of communicative English skills course at Ambo University. *East African Journal of Social Sciences and Humanities*, *3*(1), 41–62.

Immekus, J. C., Snyder, K. E., & Ralston, P. A. (2019). Multidimensional item response theory for factor structure assessment in educational psychology research. *Frontiers in Education*, *4,* 45*.* https://doi.org/10.3389/feduc.2019.00045

Karim, S. A., Sudiro, S., & Sakinah, S. (2021). Utilising test items analysis to examine the level of difficulty and discriminating power in a teacher-made test. *EduLite Journal of English Education Literature and Culture*, *6*(2), 256. https://doi.org/10.30659/e.6.2.256-269

Kasman, K., & Lubis, S. K. (2022). Teachers' performance evaluation instrument designs in the implementation of the new learning paradigm of the MerDeka curriculum. *Jurnal Kependidikan Jurnal Hasil Penelitian Dan Kajian Kepustakaan Di Bidang Pendidikan Pengajaran Dan Pembelajaran*, *8*(3), 760. https://doi.org/10.33394/jk.v8i3.5674

Kissi, P., Baidoo-Anu, D., Anane, E., & Annan-Brew, R. K. (2023). Teachers' test construction competencies in an examination-oriented educational system: Exploring teachers' multiple-choice test construction competence. *Frontiers in Education*, *8*. https://doi.org/10.3389/feduc.2023.1154592

Lee, Y. (2019). Estimating student ability and problem difficulty using item response theory (IRT) and TrueSkill. *Information Discovery and Delivery*, *47*(2), 67–75. https://doi.org/10.1108/idd-08-2018-0030

Lord, F.M. (1980). *Applications of item response theory to practical testing problems* (1st ed.). Routledge. https://doi.org/10.4324/9780203056615

Maharani, A. V., & Putro, N. H. P. S. (2020). Item analysis of the English final semester test. *Indonesian Journal of EFL and Linguistics*, *5*(2), 491. https://doi.org/10.21462/ijefl.v5i2.302

Metsämuuronen, J. (2022). Seeking the real item difficulty: Bias-corrected item difficulty and some consequences in Rasch and IRT modelling. *Behaviormetrika*, *50*(1), 121–154. https://doi.org/10.1007/s41237-022-00169-9

McTighe, J., & Ferrara, S. (2021). *Assessing student learning by design: Principles and practices for teachers and school leaders*. Teachers College Press.

Nkansah, B. K. (2018). On the Kaiser-Meier-Olkin's measure of sampling adequacy. *Mathematical theory and modeling*, *8*(7), 52-76.

Nitko, A. J., & Brookhart, S. M. (2014). *Educational assessment of students* (7th ed.). Pearson Education.

Odukoya, J. A., Adekeye, O., Igbinoba, A. O., & Afolabi, A. (2017). Item analysis of university-wide multiple choice objective examinations: The experience of a Nigerian private university. *Quality & Quantity*, *52*(3), 983–997. https://doi.org/10.1007/s11135-017-0499-2

Osterlind, S. J. (2006). *Modern measurement: Theory, principles, and applications of mental appraisal*. Pearson Education.

Pastore, S. (2023). Teacher assessment literacy: A systematic review. *Frontiers in Education*, *8*. https://doi.org/10.3389/feduc.2023.1217167

Reeve, B. (2023). Item Response Theory [IRT]. In: Maggino, F. (eds) *Encyclopedia of quality of life and well-being research*. Springer, Cham. https://doi.org/10.1007/978-3-031-17299-1_1556

Setiawati, F. A., Izzaty, R. E., & Hidayat, V. (2018). Items' parameters of the space-relations subtest using item response theory. *Data in Brief*, *19*, 1785–1793. https://doi.org/10.1016/j.dib.2018.06.061

Sharma, P. (2015). Standards-based assessments in the classroom. *Contemporary Education Dialogue*, *12*(1), 6–30. https://doi.org/10.1177/0973184914556864

Shaw, S., Crisp, V., & Johnson, N. (2012). A framework for evidencing assessment validity in large-scale, high-stakes international examinations. *Assessment in Education: Principles, Policy and Practice*, *19*(2), 159–176. https://doi.org/10.1080/0969594x.2011.563356

Sun, J. C. Y., Wu, Y. T., & Lee, W. I. (2017). The effect of the flipped classroom approach to OpenCourseWare instruction on students' self-regulation. *British Journal of Educational Technology*, *48*(3), 713-729. https://doi.org/10.1111/bjet.12444

Sundqvist, P., Wikström, P., Sandlund, E., & Nyroos, L. (2017). The teacher as examiner of L2 oral tests: A challenge to standardisation. *Language Testing*, *35*(2), 217–238. https://doi.org/10.1177/0265532217690782

Sweeney, S. M., Sinharay, S., Johnson, M. S., & Steinhauer, E. W. (2022). An investigation of the nature and consequences of the relationship between IRT difficulty and discrimination. *Educational Measurement Issues and Practice*, *41*(4), 50–67. https://doi.org/10.1111/emip.12522

Vatterott, C. (2015). *Rethinking grading: Meaningful assessment for standards-based learning*. ASCD.

Wauters, K., Desmet, P., & Van Den Noortgate, W. (2010). Adaptive item-based learning environments based on the item response theory: possibilities and challenges. *Journal of Computer Assisted Learning*, *26*(6), 549–562. https://doi.org/10.1111/j.1365-2729.2010.00368.x

William, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, *37*(1), 3–14. https://doi.org/10.1016/j.stueduc.2011.03.001

Wilson, M. (2023). *Constructing measures: An item response modelling approach* (2nd ed.). Routledge. https://doi.org/10.4324/9781003286929

Wuntu, N. V. L. E. S. C. (2021). Analysis of teacher-made tests used in summative evaluation at SMP Negeri 1 Tompaso. *Zenodo (CERN European Organisation for Nuclear Research)*. https://doi.org/10.5281/zenodo.5775342

Young, V. M., & Kim, D. H. (2010). Using assessments for instructional improvement: A literature review. *Education Policy Analysis Archives*, *18*, 19. https://doi.org/10.14507/epaa.v18n19.2010

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

Zieky, M. J. (2016). Fairness in test design and development. In: Dorans, N. J., & Cook, L. L. (Eds). *Fairness in educational assessment and measurement* (pp. 9–31). Routledge.