http://ejournal.iaingorontalo.ac.id/index.php/JETLI ISSN 2963-9565

Measuring the Quality of Teacher-Constructed English Test as Final Examination through Item Response Theory

Novri Pahrizal[1]*, Lian Gafar Otaya[2]
[1]Institut Agama Islam Negeri (IAIN) Kerinci, Jambi
[2]Institut Agama Islam Negeri (IAIN) Sultan Amai Gorontalo
pahrizal.novri@gmail.com*

Article information:
Received : 20 Juli 2025
Revised : 24 Juli 2025
Accepted : 31 Juli 2025

Abstract

This study aimed to examine the psychometric quality of a teacher-constructed English final examination test for Grade X students at Senior High School in Sungai Penuh using the framework of Item Response Theory (IRT). The analysis focused on evaluating model fit, item difficulty, and item discrimination parameters across the 1-Parameter Logistic (1-PL) and 2-Parameter Logistic (2-PL) models. Data was collected from students' responses to 40 multiple-choice items and analyzed using RStudio. The goodness-of-fit results revealed that the 2-PL model provided a better representation of the data, with 36 items classified as fit and only 3 misfitting, compared to the 1-PL model where 32 items fit and 8 misfits. Furthermore, the difficulty parameter (b) indicated that all items were within the acceptable range ($-2 \leq b \leq +2$), with a tendency toward easy to moderate levels. The discrimination parameter (a) demonstrated that most items possessed satisfactory to high discrimination power, although a small number exhibited lower values. These findings confirm that the teacher-constructed test generally meets psychometric standards of validity and reliability, while also highlighting the need for revision of a few misfitting and low-discrimination items. The study provides both theoretical and practical contributions by emphasizing the importance of applying IRT in school-based assessment practices to ensure fair, accurate, and effective evaluation of students' learning outcomes.

Keywords: teacher-constructed test, quality of test, item response theory.

INTRODUCTION

Assessment and instruction are intrinsically interconnected concepts and procedures within the educational environment, rendering their separation impractical for the attainment of high-quality and successful educational outcomes. The process of assessment can reveal invaluable insights, providing the necessary evidence to directly inform and improve the quality of teaching and learning (Reynolds, Livingston, Willson, & Willson, 2010: 2). Popham (2009) argues that high-quality education cannot be achieved without the application of effective and well-designed assessment methods. Therefore, assessment that is carefully crafted and effectively applied forms an essential foundation for achieving meaningful, effective, and high-quality learning. In language learning, assessment plays a pivotal role not only in evaluating learners' proficiency but also in providing diagnostic feedback that supports the integrated development of language skills.

Language assessment refers to an organized and standardized process of evaluation whose objective is to assess language skills/proficiencies and competencies of learners. This goes beyond taking measurements in terms of the mastery of discrete linguistic features, including lexico-grammatical knowledge and phonology, to include an evaluation of receptive (listening and reading) and productive (speaking and writing) modes. Language assessment enables the achievement of a holistic picture of the level of communication competence in a student with a well-developed methodological approach. Moreover, the obtained information can be used to diagnose personal learning requirements, guide the development of more effective teaching methods, and learn the longitudinal progress of linguistic performance. Language assessment is therefore not only a summative measure of a learning outcome as it can also be used as a diagnostic and formative tool, which is necessary to streamline the goal of language learning.

In educational practice, the final examination is strategically significant as the instrument through which comprehensive assessment of student learning outcomes can be conducted. Such summative assessment is essential for making important academic choices, such as moving up a grade, being eligible to graduate, and planning how to deliver high-quality instruction. Previous research corroborates this perspective, indicating that summative assessments yield critical evidence for high-stakes educational decision-making (Nitko & Brookhart, 2014; Popham, 2009).. In Indonesia, final exams are the most common way to test students, but their quality is

often in doubt because there aren't any systematic procedures in place, like item analysis (Mardapi, 2017). Additionally, in the domain of language education, summative assessments are considered a primary method for assessing learners' communicative competence and overall language proficiency at the conclusion of instruction (Brown & Abeywickrama, 2010).

To fully comprehend a student's abilities, it is necessary to have a good and appropriate way to measure their skills, knowledge, and learning outcomes from these tests. The principal aim of measurement is to illustrate an individual's psychological characteristics and their fluctuations (Price, 2017, p. 2). Measurement outcomes gain enhanced significance and utility when the resultant reports are intelligible across multiple dimensions; specifically, when they provide precise information for decision-makers and reduce the likelihood of misinterpretation of the results (Krisna, Mardapi, & Azwar, 2016). To show psychological traits, you need to use different types of measurement or classification systems. Measurement is primarily focused on the techniques utilized to provide quantitative indicators regarding the degree to which individuals possess or exhibit a specific attribute.

The misalignment between teacher-created test items and established learning objectives is a prevalent concern in educational assessment, significantly affecting the validity of assessments and leading to inaccurate representations of student proficiency. Empirical studies reveal that misalignment can undermine assessment validity, ultimately impairing educational decisions concerning remediation and instructional planning. For instance, Popham has pointed out that alignment between instructional materials and assessment is critical for ensuring that tests accurately reflect students' learning outcomes, and poor alignment can result in misleading interpretations of their abilities (Basuki & Anggoro, 2021). Furthermore, research by Ohiri and Okoye emphasizes the necessity of aligning assessment tasks with relevant learning objectives to bolster the validity and reliability of assessments administered (Ohiri & Okoye, 2023). This misalignment results in erroneous student proficiency data, creating negative repercussions on instructional decisions.

To address these issues effectively, integrating robust psychometric frameworks is crucial. The work of Hambleton et al. illustrates how implementing psychometric analysis enhances the quality of assessment items by ensuring they are constructed with appropriate difficulty levels and discriminating power (Sumintono, 2018). In this context, the use of

established models not only bolsters the scientific validity of assessments but also leads to more accurate ability estimates for students. For example, research has indicated that employing the Rasch model, an IRT framework, provides insights into the effectiveness and appropriateness of test items, thereby improving the quality of assessments available to educators (Sumintono, 2018). The effective construction of test items informed by psychometric principles can lead to a higher alignment with intended curricular competencies (Ohiri & Okoye, 2023).

Numerous studies employing Item Response Theory (IRT) have successfully assessed test quality, especially within large-scale, standardized assessments. These investigations have significantly contributed to our understanding of key psychometric properties, including item validity, instrument reliability, and accuracy in ability estimation (Bichi & Talib, 2018), Zanon et al., 2016). For example, Bichi and Talib emphasized how IRT frameworks facilitate the examination of individual test item characteristics, such as difficulty and discrimination, thereby enhancing overall test quality (Bichi & Talib, 2018). Similarly, Zanon et al. illustrated the advantages of IRT in evaluating the psychometric properties of educational assessments, reinforcing its efficacy in measuring constructs reliably Zanon et al., 2016). This strong foundation of research underscores the critical impact IRT has had on enhancing the measurement frameworks used in educational environments.

Despite these advancements, a notable gap exists regarding the application of IRT to assessments designed by classroom teachers, particularly those utilized as end-of-semester examinations in subjects like English language classes. Much of the current research has predominantly focused on higher-stakes assessments, including national exams and university entrance tests, which may not effectively capture the complexities of everyday classroom evaluations (Wahyuni et al., 2024), (Gavett & Horwitz, 2011;. For instance, research by Wahyuni et al. demonstrated how IRT frameworks can be effectively utilized in educational assessments, yet most applications remain confined to controlled testing environments rather than the dynamic context of classroom assessments designed by teachers (Wahyuni et al., 2024). The implications of this oversight are significant; classroom assessments often possess unique item characteristics and contextual factors that can significantly differ from those of standardized tests, leading to potentially misleading interpretations of student performance if not assessed through an appropriate IRT lens (Gavett & Horwitz, 2011; , Jahrami, 2025).

Moreover, the failure to integrate IRT practices into teacher-designed assessments can hinder the quality and effectiveness of these evaluations. Evidence indicates that IRT offers powerful tools for optimizing test design, enabling teachers to analyze the psychometric properties of their assessments more rigorously (Bósquez et al., 2025), Zanon et al., 2016). For instance, Bósquez et al. explored the psychometric properties of neuropsychological tests through an IRT approach, illustrating how this framework can effectively inform the development and validation of classroom-level assessments (Bósquez et al., 2025). Hence, there is a pressing need for educators to leverage IRT to refine their assessment practices, ensuring that items not only align with educational standards but also function effectively according to diverse student abilities.

Given that teacher-made tests play a strategic role in determining student learning outcomes and serve as a basis for pedagogical decisions—such as final grading and subsequent instructional planning—this study aims to provide a novel contribution by analyzing the quality of English tests constructed by teachers using an IRT approach. The research focuses on two key aspects: model fit measurement and the identification of item characteristics. Through this approach, the study is expected to offer a more accurate depiction of summative assessment quality while yielding practical recommendations for enhancing the quality of English language assessments at the classroom level.

METHODS

This study employs a descriptive quantitative approach to systematically evaluate the quality of teacher-constructed English tests using empirical data. The quantitative methodology was selected because the research data are numerical and require statistical analysis to derive objective information regarding the characteristics of the assessment instrument (Cohen, et.al., 2002; Creswell, 2016). Descriptive quantitative research serves to summarize, present, and explain real-world conditions numerically, thereby providing a factual representation of the quality of the analyzed test items (Bungin, 2015).

The research subjects consisted of 148 tenth-grade senior high school students in Sungai Penuh, distributed across four classes. Participant selection was based on uniformity in the instructing teachers and the identical final semester examination administered to all students. Data were collected through documentation techniques, utilizing secondary data in the form of

archival records from the final semester examination. These included exam question sheets, answer keys, student name lists, and completed student answer sheets (Widoyoko, 2014). The use of secondary data is appropriate as it provides authentic, non-reactive evidence of student performance under actual assessment conditions.

Data analysis was conducted using Item Response Theory (IRT) with the support of Rstudio software. Prior to performing model fit and item characteristic analyses, three fundamental assumptions of IRT—unidimensionality, local independence, and parameter invariance—were tested (Retnawati, 2014). The models applied included the Rasch (1-PL), 2-PL, and 3-PL models, with the best model selected based on goodness-of-fit assessed through chi-square tests and probability metrics. Subsequently, item characteristics were analyzed based on difficulty (b), discrimination (a), and guessing (c) parameters. The quality of the test items was interpreted according to the criteria established by Hambleton and Swaminathan (1985), thereby providing a comprehensive evaluation of the teacher-developed final semester examination.

RESULTS AND DISCUSSION

Results

The analysis of the final semester English examination for students of upper grades of Elementary school was conducted using Item Response Theory (IRT) with *RStudio software*. Prior to selecting an appropriate logistic model, three fundamental IRT assumptions were tested: unidimensionality, local independence, and parameter invariance.

The unidimensionality test through factor analysis revealed 13 factors with eigenvalues exceeding 1, with two dominant factors accounting for over 60% of the variance. This indicates that the test does not exhibit perfect unidimensionality, though it remains suitable for further IRT analysis.

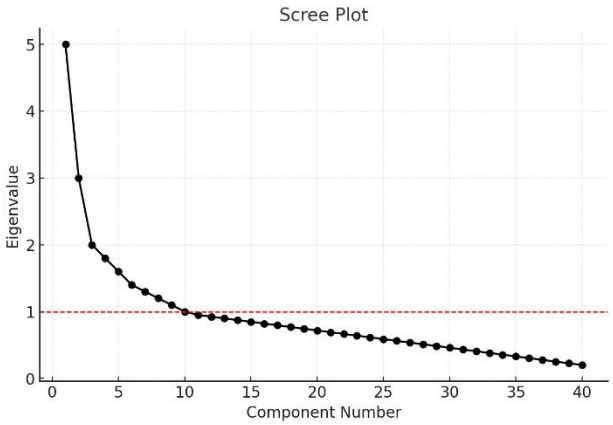

Figure 1. Scree plot

Parameter invariance in item analysis refers to the consistency of item characteristics when administered to different groups of test-takers. The two principal dimensions analyzed are item difficulty, which reflects the probability of a correct response, and discriminating power, which indicates an item's ability to differentiate between high- and low-ability examinees. When these parameters remain stable across populations, the items are considered fair, valid, and reliable for use in evaluative assessments. This stability is a fundamental prerequisite for robust measurement and meaningful score interpretation.

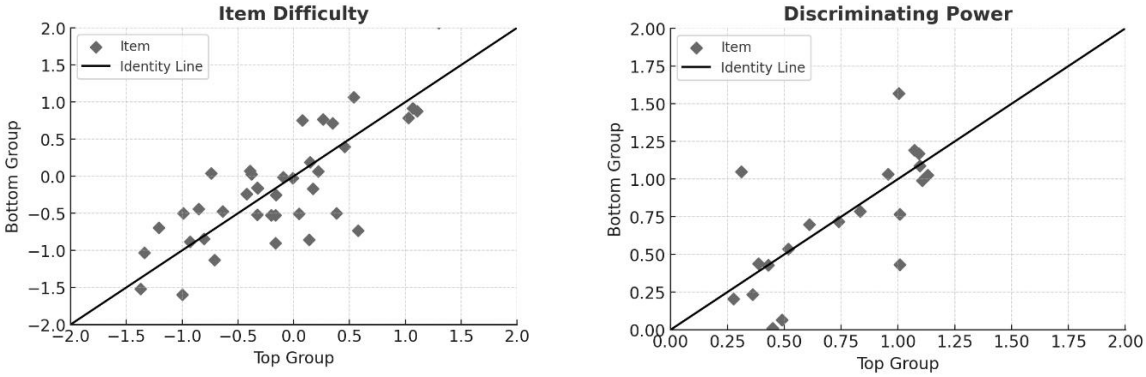

Figure 1. Plot of Item Difficulty and Discrimination Power

The first plot illustrates a comparison of item difficulty parameters between the top (high-ability) and bottom (low-ability) groups. The majority of item points cluster closely around the identity line (x = y), indicating that the difficulty levels remain relatively consistent across both groups. This suggests that the teacher-constructed English final exam demonstrates a generally fair measurement property, as no significant differential difficulty is observed between the two groups. However, a small number of items deviate noticeably from the identity line, warranting further investigation as these may exhibit potential bias or Differential Item Functioning (DIF).

The second plot displays the discrimination parameters of the items, reflecting their ability to distinguish between high- and low-ability students. Similar to the difficulty parameters, most points align near the identity line, indicating acceptable consistency in discriminatory power across groups. Nonetheless, several items show noticeable divergence, suggesting that their ability to consistently discriminate among students varies depending on the group. Such instability may reduce the items' effectiveness in measuring student ability in a fair and equitable manner.

1. Proving Model Fit of Teacher-constructed English Final Examination Test

According to established psychometric principles, evaluating the fit of items to a specified measurement model is essential for ensuring the validity and reliability of assessments. As Retnawati (2014, pp. 24–25) explains, model fit can be statistically verified through probability significance values. If the probability significance value exceeds the predetermined significance level (α), the item is classified as misfitting. Conversely, if the value is less than or equal to α, the item is considered to demonstrate adequate fit. This criterion forms the methodological basis for examining model fit in the teacher-constructed English final examination items, ensuring that the test accurately reflects the intended measurement model.

Tabel 1. The Result of Model Fit

Item	1 PL			2 PL			Item	1 PL			2 PL		
	Prob	α	Note	Prob	α	Note		Prob	α	Note	Prob	α	Note
1	0.0215	0.05	Not Fit	0.9732	0.05	Fit	22	0.0418	0.05	Not Fit	0.3481	0.05	Fit
2	0.6123	0.05	Fit	0.7594	0.05	Fit	23	0.4620	0.05	Fit	0.5377	0.05	Fit
3	0.0721	0.05	Fit	0.0932	0.05	Fit	24	0.5281	0.05	Fit	0.1625	0.05	Fit
4	0.1284	0.05	Fit	0.5221	0.05	Fit	25	0.0314	0.05	Not Fit	0.8746	0.05	Fit
5	0.0976	0.05	Fit	0.4012	0.05	Fit	26	0.0798	0.05	Fit	0.1173	0.05	Fit
6	0.0142	0.05	Not Fit	0.0915	0.05	Fit	27	0.5931	0.05	Fit	0.9672	0.05	Fit
7	0.0337	0.05	Not Fit	0.6741	0.05	Fit	28	0.6125	0.05	Fit	0.5541	0.05	Fit
8	0.1573	0.05	Fit	0.9035	0.05	Fit	29	0.9054	0.05	Fit	0.4916	0.05	Fit
9	0.0420	0.05	Not Fit	0.8124	0.05	Fit	30	0.2147	0.05	Fit	0.1349	0.05	Fit
10	0.1342	0.05	Fit	0.6317	0.05	Fit	31	0.2953	0.05	Fit	0.5410	0.05	Fit
11	0.6884	0.05	Fit	0.7210	0.05	Fit	32	0.0089	0.05	Not Fit	0.9623	0.05	Fit
12	0.0195	0.05	Not Fit	0.0029	0.05	Not Fit	33	0.9447	0.05	Fit	0.9124	0.05	Fit
13	0.0951	0.05	Fit	0.9831	0.05	Fit	34	0.2381	0.05	Fit	0.0043	0.05	Not Fit
14	0.3537	0.05	Fit	0.1958	0.05	Fit	35	0.3765	0.05	Fit	0.6941	0.05	Fit
15	0.2479	0.05	Fit	0.8920	0.05	Fit	36	0.5882	0.05	Fit	0.5436	0.05	Fit
16	0.1215	0.05	Fit	0.7542	0.05	Fit	37	0.0041	0.05	Not Fit	0.1672	0.05	Fit
17	0.7013	0.05	Fit	0.5872	0.05	Fit	38	0.1086	0.05	Fit	0.1851	0.05	Fit
18	0.7994	0.05	Fit	0.9251	0.05	Fit	39	0.8021	0.05	Fit	0.4972	0.05	Fit
19	0.9186	0.05	Fit	0.8012	0.05	Fit	40	0.1395	0.05	Fit	0.0341	0.05	Not Fit
20	0.2627	0.05	Fit	0.4331	0.05	Fit	**Fit**			**32**	**Fit**		**36**
21	0.6932	0.05	Fit	0.5783	0.05	Fit	**Not Fit**			**8**	**Not Fit**		**3**

Based on the Goodness-of-Fit analysis of the 1-Parameter Logistic (1-PL) and 2-Parameter Logistic (2-PL) models, a notable difference was observed in the number of items meeting the fit criteria. Under the 1-PL model, 32 items were classified as fitting the model, while 8 items were

deemed misfitting. In contrast, the 2-PL model demonstrated superior performance, with 36 items exhibiting satisfactory fit and only 3 items identified as misfitting. This outcome indicates that the 2-PL model more accurately represents the underlying characteristics of the data compared to the 1-PL model. The enhanced performance of the 2-PL model can be attributed to its incorporation of both item difficulty and discrimination parameters, enabling a more comprehensive capture of item variability than the 1-PL model, which estimates difficulty parameters alone.

Regarding the 3-Parameter Logistic (3-PL) model, analysis was not conducted due to sample size limitations and potential instability in estimating the guessing parameter within this dataset. While the 3-PL model is theoretically more complex—incorporating difficulty, discrimination, and guessing parameters—it requires a relatively larger sample size and more diverse data distribution to achieve accurate parameter estimation. Given the constraints of the current dataset, applying the 3-PL model could have led to non-convergent or biased estimates. Consequently, the analysis was restricted to the 1-PL and 2-PL models, which align more appropriately with the data characteristics and ensure methodological robustness (de Ayala, 2009; Hambleton & Swaminathan, 1985).

2. Proving Item Characteristic of Teacher-Constructed English Final Examination Test

Based on the results of the model fit analysis, further examination of the item characteristics in the teacher-constructed English Final Semester Examination was conducted using the Two-Parameter Logistic (2-PL) model. The selection of this model was grounded in previous findings, which demonstrated that the 2-PL model exhibited a better fit compared to the 1-PL model, as reflected in the higher number of items satisfying the fit criteria. The 2-PL model is considered more representative due to its ability to capture two essential aspects of each item: the difficulty parameter and the discrimination parameter. This allows for a more comprehensive interpretation of the item characteristics, enhancing the validity and reliability of the test evaluation.

Table 2. Results of Item Difficulty for Teacher-constructed English Final Examination Test

ITEMS	ITEM DIFFICULTY/ INTERCEPT	α	Note	ITEMS	ITEM DIFFICULTY/ INTERCEPT	α	Note
1	-1.211	(-2 ≤ b ≤ 2)	Fit	22	-0.571	(-2 ≤ b ≤ 2)	Fit
2	-0.169	(-2 ≤ b ≤ 2)	Fit	23	-0.587	(-2 ≤ b ≤ 2)	Fit
3	-0.363	(-2 ≤ b ≤ 2)	Fit	24	-0.768	(-2 ≤ b ≤ 2)	Fit
4	-0.557	(-2 ≤ b ≤ 2)	Fit	25	-0.401	(-2 ≤ b ≤ 2)	Fit
5	-0.471	(-2 ≤ b ≤ 2)	Fit	26	-0.727	(-2 ≤ b ≤ 2)	Fit

6	-0.262	(-2 ≤ b ≤ 2)	Fit	27	-0.738	(-2 ≤ b ≤ 2)	Fit
7	-0.767	(-2 ≤ b ≤ 2)	Fit	28	-0.267	(-2 ≤ b ≤ 2)	Fit
8	-0.6	(-2 ≤ b ≤ 2)	Fit	29	-0.415	(-2 ≤ b ≤ 2)	Fit
9	-1.289	(-2 ≤ b ≤ 2)	Fit	30	-0.777	(-2 ≤ b ≤ 2)	Fit
10	-1.275	(-2 ≤ b ≤ 2)	Fit	31	-0.528	(-2 ≤ b ≤ 2)	Fit
11	0.037	(-2 ≤ b ≤ 2)	Fit	32	-1.148	(-2 ≤ b ≤ 2)	Fit
12	-0.299	(-2 ≤ b ≤ 2)	Fit	33	0.02	(-2 ≤ b ≤ 2)	Fit
13	-0.215	(-2 ≤ b ≤ 2)	Fit	34	0.295	(-2 ≤ b ≤ 2)	Fit
14	-0.286	(-2 ≤ b ≤ 2)	Fit	35	-0.336	(-2 ≤ b ≤ 2)	Fit
15	0.323	(-2 ≤ b ≤ 2)	Fit	36	-0.399	(-2 ≤ b ≤ 2)	Fit
16	-0.048	(-2 ≤ b ≤ 2)	Fit	37	-0.373	(-2 ≤ b ≤ 2)	Fit
17	-0.161	(-2 ≤ b ≤ 2)	Fit	38	-0.596	(-2 ≤ b ≤ 2)	Fit
18	-0.99	(-2 ≤ b ≤ 2)	Fit	39	-0.214	(-2 ≤ b ≤ 2)	Fit
19	-0.449	(-2 ≤ b ≤ 2)	Fit	40	-0.394	(-2 ≤ b ≤ 2)	Fit
20	-0.572	(-2 ≤ b ≤ 2)	Fit	**Average**	**-0.464142857**		**Fit**
21	-0.123	(-2 ≤ b ≤ 2)	Fit				

Based on the analysis of the item difficulty parameters (intercept) for the 40 items in the teacher-constructed English Final Examination, all items demonstrated values within the range of $-2 \leq b \leq +2$. Consequently, it can be concluded that all items are deemed fit according to the criteria established by Retnawati (2014). This indicates that no items were excessively easy ($b < -2$) or overly difficult ($b > +2$) for the test-takers. Thus, the overall quality of the items can be considered satisfactory in terms of difficulty level.

Furthermore, the average item difficulty was –0.464, suggesting that the test tends to fall within the easy to moderate range. This condition implies that the test primarily measures students' mastery of basic competencies, thereby providing greater opportunity for test-takers to demonstrate their abilities. Overall, these results confirm that the teacher-constructed test meets the adequacy standards in terms of item difficulty appropriateness.

Table 3. Results of Discrimination Power for Teacher-constructed English Final Examination Test

ITEMS	DESCRIMINATING POWER/ SLOPE	a	Note	ITEMS	DESCRIMINATING POWER/ SLOPE	a	Note
1	0.984	(0 ≤ a ≤ 2)	Fit	22	0.809	(0 ≤ a ≤ 2)	Fit
2	0.54	(0 ≤ a ≤ 2)	Fit	23	0.316	(0 ≤ a ≤ 2)	Fit
3	0.466	(0 ≤ a ≤ 2)	Fit	24	0.383	(0 ≤ a ≤ 2)	Fit
4	0.566	(0 ≤ a ≤ 2)	Fit	25	0.569	(0 ≤ a ≤ 2)	Fit
5	0.702	(0 ≤ a ≤ 2)	Fit	26	0.329	(0 ≤ a ≤ 2)	Fit
6	0.249	(0 ≤ a ≤ 2)	Fit	27	0.367	(0 ≤ a ≤ 2)	Fit
7	0.751	(0 ≤ a ≤ 2)	Fit	28	0.301	(0 ≤ a ≤ 2)	Fit

8	0.638	(0 ≤ a ≤ 2)	Fit	29	0.404	(0 ≤ a ≤ 2)	Fit
9	0.291	(0 ≤ a ≤ 2)	Fit	30	0.488	(0 ≤ a ≤ 2)	Fit
10	0.295	(0 ≤ a ≤ 2)	Fit	31	0.637	(0 ≤ a ≤ 2)	Fit
11	0.313	(0 ≤ a ≤ 2)	Fit	32	0.882	(0 ≤ a ≤ 2)	Fit
12	0.25	(0 ≤ a ≤ 2)	Fit	33	0.327	(0 ≤ a ≤ 2)	Fit
13	0.976	(0 ≤ a ≤ 2)	Fit	34	0.351	(0 ≤ a ≤ 2)	Fit
14	0.497	(0 ≤ a ≤ 2)	Fit	35	0.243	(0 ≤ a ≤ 2)	Fit
15	0.621	(0 ≤ a ≤ 2)	Fit	36	0.43	(0 ≤ a ≤ 2)	Fit
16	0.663	(0 ≤ a ≤ 2)	Fit	37	0.238	(0 ≤ a ≤ 2)	Fit
17	0.365	(0 ≤ a ≤ 2)	Fit	38	0.356	(0 ≤ a ≤ 2)	Fit
18	0.301	(0 ≤ a ≤ 2)	Fit	39	0.325	(0 ≤ a ≤ 2)	Fit
19	0.3	(0 ≤ a ≤ 2)	Fit	40	0.257	(0 ≤ a ≤ 2)	Fit
20	0.238	(0 ≤ a ≤ 2)	Fit	Average	0.458325		Fit
21	0.315	(0 ≤ a ≤ 2)	Fit				

Based on the analysis of the item discrimination parameters (slope, *a*) for the 40 items in the teacher-constructed English Final Examination, all items demonstrated values within the range of $0 \leq a \leq 2$ and were thus classified as psychometrically fit. This indicates that each item effectively discriminates between high- and low-ability test-takers, albeit with varying degrees of discriminatory strength. The highest discrimination values were observed for Item 1 (0.984) and Item 13 (0.976), suggesting these items are particularly effective in differentiating examinee proficiency. In contrast, Items 20 and 37 exhibited relatively lower discrimination indices (both 0.238), though they remained within acceptable fit thresholds.

The overall average discrimination power was 0.458, placing the test in the moderately effective range. These results indicate that the teacher-developed examination generally demonstrates adequate discriminatory validity, though items with lower discrimination values may benefit from revision or refinement to further enhance the measurement precision of the test (Hambleton et al., 1991).

Discussion

The results of the unidimensionality test via factor analysis revealed 13 factors with eigenvalues > 1, with two dominant factors accounting for over 60% of the variance. This finding indicates that the test is not strictly unidimensional, though it remains practically suitable for IRT analysis. This aligns with Reckase's (2009) assertion that, in measurement practice, the unidimensionality assumption need not be absolute, provided a single dominant factor sufficiently explains a substantial portion of the variance. Thus, the teacher-constructed English

Final Examination meets the basic requirements for further analysis, though its unidimensionality could be strengthened through more structured item development.

Analysis of parameter invariance also yielded relatively positive results. In the item difficulty comparison plot, the majority of items clustered near the identity line (x = y), indicating consistency in difficulty levels between high- and low-ability student groups. This suggests the test is generally fair, consistent with IRT parameter invariance theory, which emphasizes the importance of stable item characteristics (Hambleton, Swaminathan, & Rogers, 1991). However, the slight deviation of some items from the identity line warrants attention, as it may indicate potential Differential Item Functioning (DIF). This finding aligns with prior studies (Retnawati, 2016), which note that while most teacher-constructed items tend to exhibit fit, some may still display bias requiring revision or further investigation.

Regarding discriminating power, most items also demonstrated consistency across groups, indicating that their ability to distinguish between high- and low-ability students remained relatively stable. Nevertheless, the deviation of a few items suggests instability in their discriminatory effectiveness, reducing their utility as precise measures of ability. This finding partially contradicts Baker's (2001) emphasis on stable discrimination power as a key determinant of test validity. Therefore, although the test is generally adequate, items exhibiting inconsistent difficulty or discrimination levels should be reviewed to enhance measurement quality.

Overall, the results indicate that the teacher-constructed English Final Examination largely satisfies the basic IRT assumptions of unidimensionality and parameter invariance, though certain items require refinement. This aligns with previous studies (Mislevy, 1996; Retnawati, 2014), which emphasize that teacher-made tests often necessitate further evaluation and calibration to ensure they function as valid, reliable, and fair measurement tools for all students. Thus, this study underscores the importance of empirical item analysis, moving beyond content validation to incorporate comprehensive psychometric evaluation.

1. Proving Model Fit of Teacher-Constructed English Final Examination Test

The Goodness-of-Fit analysis indicates that the Two-Parameter Logistic (2-PL) model yields superior results compared to the One-Parameter Logistic (1-PL) model. Under the 1-PL model evaluation, only 32 items met the fit criteria, while 8 items were classified as misfitting. In contrast, the 2-PL model demonstrated more optimal performance, with 36 items exhibiting satisfactory fit and only 3 items identified as misfitting. This confirms the enhanced capability of

the 2-PL model in parameter estimation, as it comprehensively accounts for both essential item characteristics: difficulty and discrimination parameters (Maydeu-Olivares, 2013). This finding aligns with the arguments posited by Hambleton and Swaminathan, as well as Baker, suggesting that the 2-PL model is generally more representative in capturing the characteristics of teacher-constructed tests, offering greater flexibility in accommodating variability in test-taker responses (Maydeu-Olivares, 2013).

Empirically, studies by Retnawati (2016), Zanon et al. (2016), and Kang and Chen (2010) have similarly observed that the 2-PL model exhibits a higher proportion of fitting items compared to the 1-PL model, particularly in tests featuring a large number of items and high heterogeneity in student responses. Nevertheless, the presence of misfitting items across both models underscores the need for improvements in item construction, linguistic clarity, and alignment with learning indicators. This observation resonates with Maydeu-Olivares' perspective, which asserts that model-misfitting items can compromise measurement validity and may reflect inconsistencies in students' interpretation of item content (Maydeu-Olivares, 2013).

Within this context, continuous evaluation of misfitting items remains essential for enhancing the overall quality of the assessment instrument. These findings reinforce the argument that the 2-PL model is more appropriate for analyzing the quality of teacher-constructed tests, while simultaneously highlighting the necessity for ongoing efforts to refine item development processes and improve instructional clarity.

2. Proving Item Characteristic of Teacher-Constructed English Final Examination Test

The findings of this study reveal that all 40 items from the teacher-constructed English Final Semester Examination fall within the acceptable difficulty range of $-2 \leq b \leq +2$, indicating that none of the items are excessively easy or difficult. This result suggests that the items exhibit proportional measurement of student competencies, aligning with the principles of Item Response Theory (IRT) Uyigue & Orheruata (2019). According to Arifin and Yusoff, items within this difficulty range possess good quality and contribute positively to the overall validity of the test Arifin & Yusoff, 2017). The average difficulty value of -0.464 indicates a tendency toward easy to moderate items, reflecting a focus on assessing foundational competencies. Such a trend can be beneficial, particularly in educational contexts where ensuring base level mastery is essential Arifin & Yusoff, 2017).

However, it is critical to note that a test dominated by easier items may lack the challenge necessary for high-ability students, thereby potentially limiting its discriminative capacity between high and low performers Arifin & Yusoff, 2017). Despite all items demonstrating discrimination values within an acceptable range ($0 \leq a \leq 2$), considerable variation in discrimination power was observed among individual items. For example, high-discrimination items such as Item 1 (0.984) and Item 13 (0.976) effectively distinguish student abilities, while Items 20 and 37 showcased relatively low discrimination ($a = 0.238$) Arifin & Yusoff, 2017). These findings resonate with Retnawati's emphasis on revising items with low discrimination by enhancing clarity and alignment, which is crucial in achieving accurate diagnostics of student performance Arifin & Yusoff, 2017).

When contextualized within previous research, these results parallel observations that teacher-constructed tests typically emphasize moderate to easy difficulty levels, reflecting instructors' focus on student pass rates. Furthermore, it has been noted that while many teacher-generated items meet difficulty criteria, their discrimination capacity often warrants further attention (Cao et al., 2014; Arifin & Yusoff, 2017). While this study confirms all items meet the necessary difficulty standards, the limited discriminative ability of certain items suggests that the effectiveness of the assessment could be improved.

In comparison to international standards for effective test construction, an optimal assessment should showcase not just proportional difficulty but also strong discrimination to maximize insights into student abilities (Stone & Zhang, 2003). Although the test's average discrimination value of 0.458 is acceptable, it highlights the need for revisions, particularly for low-discrimination items to enhance measurement precision. Notably, these findings underscore the importance and relevance of modern evaluation methods, particularly IRT. The preference for the 2-PL model over the 1-PL model in this study allowed for a nuanced representation of item characteristics, encompassing both difficulty and discrimination parameters effectively. Thus, the implications of this study advocate that teachers should complement Classical Test Theory with IRT-based analyses to elevate the quality and validity of their assessment tools (Curtis, 2010).

CONCLUSIONS

Based on the analysis employing Item Response Theory (IRT), it can be concluded that the teacher-constructed English Final Semester Examination for for upper grades of Elementary school generally meets the criteria for sound test quality. Goodness-of-Fit testing revealed that

the Two-Parameter Logistic (2-PL) model was more representative than the 1-PL model, with a higher number of fitting items. All test items fell within an appropriate difficulty range ($-2 \leq b \leq +2$), with an overall tendency toward easy to moderate difficulty levels, confirming the test's emphasis on assessing mastery of basic competencies. In terms of discrimination power, all items were also deemed fit, although notable variation was observed. Some items exhibited strong discriminatory ability, while others demonstrated relatively lower values. These findings indicate that the teacher-made test is generally fair and reliable in measuring student ability. However, revisions to items with low discrimination power and enhancements to the test's unidimensionality are recommended to further optimize its measurement quality.

REFERENCES

Arifin, W. and Yusoff, M. (2017). Item response theory for medical educationists. Education in Medicine Journal, 9(3), 69-81. https://doi.org/10.21315/eimj2017.9.3.8

Baker, F. B. (2001). The Basics of Item Response Theory (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation

Basuki, L. and Anggoro, S. (2021). Improving the competency to construct test items for class vi teachers through workshop.. https://doi.org/10.4108/eai.19-7-2021.2312716

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

Brown, H. D., & Abeywickrama, P. (2010). Language assessment: Principles and classroom practices (2nd ed.). Pearson.

Brown, H. Douglas, Abeywickrama, Priyanvada. (2010). *Language Assessment : Principles and Classroom Practices* (3). New York: Pearson Education.

Cao, Y., Lu, R., & Wei, T. (2014). Effect of item response theory (irt) model selection on testlet-based test equating. Ets Research Report Series, 2014(2), 1-13. https://doi.org/10.1002/ets2.12017

Cohen, L., Manion, L., & Morrison, K. (2002). *Research methods in education*. routledge.

Creswell, J. W., & Creswell, J. D. (2016). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Curtis, S. (2010). bugscode for item response theory. Journal of Statistical Software, 36(Code Snippet 1). https://doi.org/10.18637/jss.v036.c01

de la Torre, J. (2009). Improving the quality of ability estimates through multidimensional scoring and incorporation of ancillary variables. *Applied Psychological Measurement, 33*(6), 465-485.

Gavett, B. and Horwitz, J. (2011). Immediate list recall as a measure of short-term episodic memory: insights from the serial position effect and item response theory. Archives of Clinical Neuropsychology, 27(2), 125-135. https://doi.org/10.1093/arclin/acr104

Hambleton R.K. & Swaminathan H. (1985). Items Response Theory: Principles and Application. Kluwer-Nijjhoff Publish.

Hambleton, Ronald K; Swaminathan, H; dan Jane Rogers, H. 1991. Fundamentals of Item Response Theory. London: SagePublications

Jahrami, H. (2025). The validation of the nomophobia questionnaire using a modern psychometric approach: an item response theory analysis of 5087 participants. Brain and Behavior, 15(6). https://doi.org/10.1002/brb3.70622

Kang, T. and Chen, T. (2010). Performance of the generalized s-x2 item fit index for the graded response model. Asia Pacific Education Review, 12(1), 89-96. https://doi.org/10.1007/s12564-010-9082-4

Krisna, I. I., Mardapi, D., & Azwar, S. (2016). Determining standard of academic potential based on the Indonesian Scholastic Aptitude Test (TBS) benchmark. *REID (Research and Evaluation in Education)*, *2*(2), 5.

Mardapi, D. (2017). Pengukuran, penilaian, dan evaluasi pendidikan. Yogyakarta: Parama Publishing.

Mardapi, D. (2017). Pengukuran, Penilaian, dan Evaluasi Pendidikan (Edisi 2). *Yogyakarta: Parama Publishing*.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. Measurement Interdisciplinary Research and Perspectives, 11(3), 71-101. https://doi.org/10.1080/15366367.2013.831680

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement*, *33*(4), 379-416.

Nitko, A. J., & Brookhart, S. M. (2014). Educational assessment of students (6 th ed.). Pearson Education, Inc

Ohiri, S. and Okoye, R. (2023). Application of classical test theory as linear modeling to test item development and analysis. International Research Journal of Modernization in Engineering Technology and Science. https://doi.org/10.56726/irjmets45379

Popham, W. J. (2009). Classroom assessment: What teachers need to know (6th ed.). Pearson.

Price, L. R. (2017). *Psychometric methods: Theory into practice*. New York: Guilford Publications.

Reckase, M. D. (2009). Multidimensional item response theory models *Multidimensional Item Response Theory* (pp. 79-112): Springer.

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

Retnawati, H. (2016). Analisis kuantitatif instrumen penelitian (panduan peneliti, mahasiswa, dan psikometrian). Parama publishing.

Reynolds, C. R., Livingston, R. B., Willson, V. L., & Willson, V. (2010). Measurement and assessment in education: Pearson Education International Upper Saddle River.

Stone, C. and Zhang, B. (2003). Assessing goodness of fit of item response theory models: a comparison of traditional and alternative procedures. Journal of Educational Measurement, 40(4), 331-352. https://doi.org/10.1111/j.1745-3984.2003.tb01150.x

Sumintono, B. (2018). Rasch model measurements as tools in assesment for learning.. https://doi.org/10.2991/icei-17.2018.11

Uyigue, V. and Orheruata, M. (2019). Test length and sample size for item-difficulty parameter estimation in item response theory. JEP. https://doi.org/10.7176/jep/10-30-08

Wahyuni, L., Sarwanto, S., & Atmojo, I. (2024). Measurement of science literacy skills of elementary school teacher education students: development and validity testing of assessment instruments. International Journal of Current Science Research and Review, 07(11). https://doi.org/10.47191/ijcsrr/v7-i11-27

Widoyoko (2014). Teknik Penyusunan Instrumen Penelitian. Yogyakarta: Pustaka Pelajar.

Zanon, C., Hutz, C., Yoo, H., & Hambleton, R. (2016). An application of item response theory to psychological test development. Psicologia Reflexão E Crítica, 29(1). https://doi.org/10.1186/s41155-016-0040-x