Jurnal Ilmiah Pendidikan Fisika
https://ppjp.ulm.ac.id/journals/index.php/jipf/index
e-ISSN: 2549-9963 p-ISSN: 2549-9955
Vol 10 No 2 2026
DOI: https://doi.org/10.20527/jipf.v10i2.17399

Item Analysis of Formative Tests on Physics Problem Solving in the Area of Dynamic Electricity in High School

Candra Purnama, Kaharuddin Arafah, and Suryadi Ishak*
Master of Educational Research and Evaluation Study Program,
Makassar State University, Makassar, Indonesia
*suryadi.ishak@unm.ac.id

Article Info

Article history:
Received: May 6, 2026
Accepted: July 1, 2026
Published: July 28, 2026

Keywords:
Formative assessment
IRT analysis
Problem solving

This article is licensed under a Creative Commons Attribution-Share Alike 4.0 International license

© 2026 Jurnal ilmiah Pendidikan Fisika

Abstract

The implementation and development of formative test items often still face various challenges that impact the accuracy of evaluation results. This study aims to analyze the quality of formative test items on physics problem-solving, particularly in the topic of dynamic electricity, by applying the Item Response Theory (IRT) approach in senior high schools in Bulukumba Regency. The study focuses on four main aspects: validity, reliability, item difficulty, and item discrimination. This research employed a quantitative descriptive design with data consisting of test responses from 369 twelfth-grade students. The analysis was conducted in three stages: content validity testing using Aiken's V to assess the alignment of items with the measured competencies; reliability testing using the Kuder–Richardson 20 (KR-20) coefficient; and item characteristic analysis based on the IRT approach to obtain indices of item difficulty and discrimination. The results indicate that all items demonstrated good content validity, with Aiken's V values above 0.75. The reliability coefficient showed high internal consistency, indicating that the instrument is reliable as a measurement tool. Regarding item difficulty, the 15 test items ranged from easy to difficult, with the majority categorized as moderate (10 items), while a smaller number were classified as easy (2 items) and difficult (3 items). This distribution suggests that the test adequately represents students' abilities, particularly at the average ability level. The discrimination index ranged from moderate to high, with five items categorized as high and ten items as moderate. No items fell into the low category, indicating that the test items effectively distinguish students based on their ability levels. Overall, the instrument is considered valid and reliable, with proportional levels of difficulty and good discrimination power, making it appropriate for measuring physics problem-solving competence, particularly in the topic of dynamic electricity. The implication of this study is the need for wider application of IRT-based item analysis in research on the development of educational instruments at the high school level.

To cite this article: Purnama, C., Arafah, K., & Ishak, S. **(**2026). Item Analysis of Formative Tests on Physics Problem Solving in the Area of Dynamic Electricity in High School. *Jurnal Ilmiah Pendidikan Fisika*, *10* (2), 278-292.

INTRODUCTION

Science education, particularly physics, plays a fundamental role in developing students' logical, analytical, and critical thinking skills. Physics is not merely viewed as a collection of mathematical formulas but rather as an intellectual vehicle that requires students to integrate conceptual understanding with practical application in problem-solving. One prominent topic in this context is dynamic electricity, which is often perceived as a

challenging subject by students. This is due to the subject's connection to abstract concepts, the use of complex mathematical representations, and its relevance to everyday life situations. Thus, mastery of dynamic electricity not only tests students' cognitive abilities but also serves as a means to develop the higher-order thinking skills expected in modern science curricula.

Assessment plays a central role as a tool for identifying the extent to which students understand the material and have internalized problem-solving skills. Formative tests (Andrade et al., 1988; Andrade & Heritage, 2018), as part of the evaluation process, serve not only as a measure of learning achievement but also as a diagnostic tool that provides constructive feedback to teachers in designing subsequent learning strategies. However, the significance of formative tests is largely determined by the quality of the test items developed. Good test items must be able to represent the depth of concepts, foster critical thinking skills, and connect theory with real-world contexts so that they truly measure essential learning outcomes. Therefore, the development of valid, reliable, and problem-solving-oriented assessment instruments has become an urgent need in physics education

The assessment instruments used in physics education often fail to meet ideal measurement quality standards. Aspects such as content validity, reliability, discriminability, and item difficulty are frequently overlooked during the development process, such as in the topics of mechanics and energy. A high-quality test must measure what it is intended to measure (validity) (Beck, 2020), provide consistent results (reliability) (Azwar, 2016), be able to distinguish between high- and low-ability students (discriminative power), and have a balanced level of difficulty (Arikunto, 2019). In line with Nitko, high-quality assessment instruments will yield accurate information regarding students' abilities (Nitko & Brookhart, 2011). Thus, if these aspects are neglected, formative test results may fail to reflect students' actual competencies, thereby undermining the evaluative function in supporting the achievement of more meaningful learning objectives.

One strategy for ensuring the quality of an assessment instrument is to conduct item analysis. The quality of an assessment instrument not only determines the success of the learning process but also serves as a key indicator of the achievement of broader educational goals. Research on item analysis using the Item Response Theory (IRT) approach (Magnus, 2018) has been widely conducted, but most studies remain focused on summative assessment contexts, such as final exams or concept inventory tests. For example, the study conducted by Syafii et al. (2021) analyzed the items from the End-of-Year Assessment in the subject of Al-Qur'an and Hadith using a two-parameter IRT model. The results of this study showed a relatively balanced distribution of item difficulty and discrimination. Although it contributes to improving the quality of summative evaluation, this study has not been directed toward the development of IRT-based instruments in the realm of formative assessment, which actually plays a crucial role in continuously monitoring students' learning progress.

To address this gap, the researcher conducted an item analysis of a formative assessment instrument for physics problem-solving in the area of dynamic electricity using the IRT 2PL model. This study aims to: 1) examine the content validity and reliability of the formative test instrument, 2) analyze the difficulty level of the items in the instrument, and 3) analyze the discriminant power of the items in the physics problem-solving instrument on the topic of dynamic electricity. The results of the analysis are expected to serve as a reference for the development of formative test instruments, particularly in the subject of physics.

METHODS

This study employed a descriptive quantitative approach using item analysis of formative physics problem-solving tests based on the IRT model, specifically the two-parameter logistic (2PL) model. This study falls under the category of quantitative descriptive research, as it aims to describe the quality of physics problem-solving formative test items based on IRT analysis results. The items were first developed based on indicators of physics problem-solving ability, followed by the administration of a formative test to the students. The test results are then analyzed using IRT 2PL-based software, which generates parameter values a and b for each test item. These results are compared with item quality criteria to determine the suitability of each item. Items that meet the validity and reliability criteria are retained, while problematic items are revised or removed from the item bank. The formative test in this study consists of multiple-choice questions with five answer choices (A, B, C, D, E) and a total of 15 items. Figure 1 shows some sample questions that have been prepared.

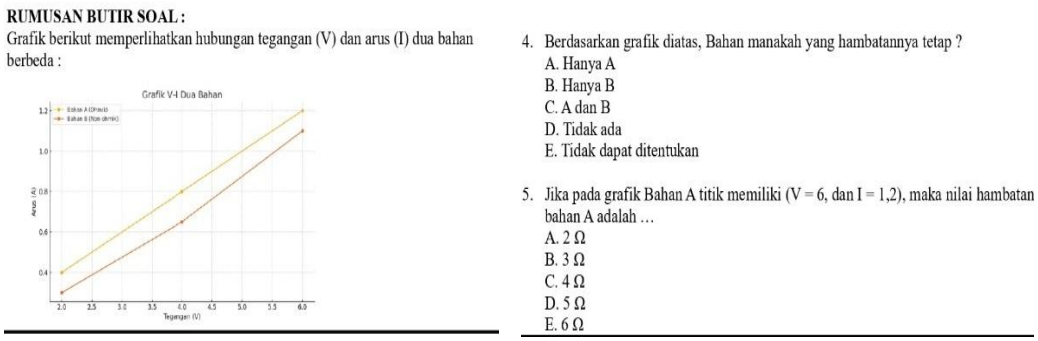

Figure 1 Examples of items 4 and 5

The subjects in this study were students who had studied the topic of dynamic electricity in their physics course. The subjects were selected because they were the direct users of the formative test developed to measure physics problem-solving skills. The number of subjects was adjusted to meet the requirements of IRT analysis, specifically a minimum of 200 respondents (Linacre, 1994), to ensure more accurate item parameter estimates. The research participants came from a class at the high school level during the even semester of the current academic year. The test results from these subjects were used as the primary data to analyze item characteristics, including difficulty and discriminant power, using the IRT model. A total of 369 students were analyzed in this study.

This study employed the documentation technique as the primary method of data collection. The data collected included the test items for formative assessment of physics problem-solving on the topic of dynamic electricity, students' responses to those items, the answer keys used in grading, and the item development guidelines. All documents were systematically collected from the test administration in the classroom and then converted into digital formats, such as Microsoft Word and Excel files, to facilitate the analysis process using R software version 4.4.2. The main instruments analyzed in this study include physics formative test questions, student answer sheets, answer keys, and question guidelines developed based on problem-solving ability indicators.

The process of analyzing the collected data, with the aim of obtaining information that is relevant and consistent with the established focus and objectives of the study. Data analysis in this study was conducted through several stages, namely: 1. Content validity and reliability testing, 2. Item characteristic analysis, 3. Prerequisite testing, 4. Sample

adequacy testing, 5. Unidimensionality testing, 6. Local independence testing, 7. Model fit testing, 8. Item parameter estimation, 9. Test item quality evaluation.

Content validation of the physics problem-solving items on the topic of dynamic electricity was conducted by involving experts who rated each item on a scale of 1 to 4. The evaluation covered five main aspects: scientific competence, subject knowledge, item context, spelling and language, and clarity of instructions. After content validation was completed, the next step was to analyze the item characteristics using the IRT approach with a two-parameter logistic (2PL) model. This analysis was performed using R software version 4.4.2, which yielded two primary parameters: discriminant power, indicating the item's ability to distinguish between high- and low-ability students, and difficulty level, indicating the level of ability required to answer correctly with a 50% probability. Before conducting item analysis using the IRT 2PL model (Ankomah & Nugba, 2020), a series of prerequisite tests must be performed to ensure that the applied model aligns with the characteristics of the analyzed data. There are three main prerequisites that must be met in the application of the 2PL IRT model, namely the test of sample adequacy, unidimensionality, and local independence. Fulfilling these three prerequisites is important to ensure the validity of item parameter estimates and student ability. To ensure that the 2PL model used is appropriate for the student response data, a goodness-of-fit test was conducted.

A quantitative approach was used because the data analyzed consisted of students' test scores, which were statistically processed to obtain item parameters. The research design employed was an item analysis design, which focuses on testing the characteristics of test items to determine their validity, reliability, difficulty (b), and discrimination (a). The research procedure was carried out through the stages of test instrument development, test administration to respondents, collection of test results, and data analysis using IRT (2PL) models. With this design, the researcher can objectively describe the quality of each test item in measuring students' problem-solving abilities in the subject of dynamic electricity.

RESULTS AND DISCUSSION

Validity and Reliability Testing

A good test must be able to measure what it is intended to measure (validity) and produce consistent results (reliability) (Arikunto, 2019). Content validity was estimated using the V-Aiken's method, while the reliability of the instrument was estimated using the KR20 approach. The result of content validity using Aiken's V is shown in Table 1.

Table 1 Content validity using Aiken's V

Item	Aiken's V Average	Aiken's V Critical Value	Description
1	0.86	0.75	Valid
2	0.88	0.75	Valid
3	0.83	0.75	Valid
4	0.81	0.75	Valid
5	0.85	0.75	Valid
6	0.87	0.75	Valid
7	0.82	0.75	Valid
8	0.90	0.75	Valid
9	0.84	0.75	Valid
10	0.86	0.75	Valid
11	0.80	0.75	Valid
12	0.83	0.75	Valid
13	0.88	0.75	Valid
14	0.85	0.75	Valid
15	0.87	0.75	Valid

Based on the results of the content validity analysis using Aiken's V coefficient (Aiken, 1985), Table 1 presents the content validity using Aiken's V, which involved five validators assessing 11 evaluation criteria; the validity scores for each item ranged from 0.80 to 0.90. All items had an Aiken's V value greater than the established critical value of 0.75; therefore, it can be concluded that the instrument possesses excellent content validity and is suitable for use in research.

The results of the reliability analysis using the Kuder–Richardson 20 (KR-20) coefficient yielded a value of 0.84. This value indicates that the Physics Formative Test has a level of reliability that is considered good. This coefficient indicates a sufficiently strong internal consistency among the test items, in accordance with general reliability criteria stating that a coefficient value of 0.70 or higher meets the acceptability threshold for various forms of measurement (Salkind et al., 2020).

Prerequisite Test

Sample Suitability Tests

There are three main criteria that must be met when applying the IRT 2PL model: 1) the Kaiser-Meyer-Olkin (KMO)-Bartlett test to assess sample adequacy, 2) unidimensionality, and 3) local independence. Meeting these three criteria is essential to ensure the validity of item parameter estimates and student learning outcomes. The results of the KMO and Bartlett tests are shown in Table 2.

Table 2 KMO and Bartlett's tests

No.	Statistic	Value
1	KMO Overall	0.842
2	Bartlett Chi-Square	1287.456
3	Bartlett df	105
4	Bartlett p-value	0.000

Table 2 shows that the results of the KMO test showed a value of 0.842, which falls into the "good" category; therefore, the sample size was deemed adequate for factor analysis. Furthermore, the results of Bartlett's test of sphericity showed a chi-square value of 1287.456 with 105 degrees of freedom and a significance of 0.000 ($p < 0.05$). These results indicate that the inter-item correlation matrix is not an identity matrix, so the data are suitable for further analysis using factor analysis.

Unidimensionality

The assumption of unidimensionality in IRT analysis refers to the premise that students' responses to all test items are influenced by a single dominant latent factor, namely physics problem-solving ability. In other words, all test items should measure the same construct. The assumption of unidimensionality is considered to be met if the eigenvalue of the first component is significantly larger than the eigenvalues of the subsequent components. The results of the unidimensionality test are shown in Table 3.

Table 3 Test for unidimensionality

Factor	Eigenvalue	PropVar (%)	CumVar (%)
1	6.72	44.80	44.80
2	0.98	6.53	51.33
3	0.87	5.80	57.13
4	0.79	5.27	62.40
5	0.73	4.87	67.27
6	0.69	4.60	71.87
7	0.63	4.20	76.07
8	0.59	3.93	80.00

Factor	Eigenvalue	PropVar (%)	CumVar (%)
9	0.55	3.67	83.67
10	0.50	3.33	87.00
11	0.46	3.07	90.07
12	0.41	2.73	92.80
13	0.36	2.40	95.20
14	0.32	2.13	97.33
15	0.40	2.67	100.00

Table 3 shows that there is only one component with an eigenvalue greater than 1, specifically 6.72, accounting for 44.80% of the variance. Meanwhile, the other components have eigenvalues less than 1.

In addition to using the eigenvalue approach, another method for testing unidimensionality involves scree plot analysis. To determine the number of factors using a scree plot, one must carefully observe the shape of the slope. Therefore, the scree plot approach requires a fairly high degree of precision. Scree plot is shown in Figure 2.

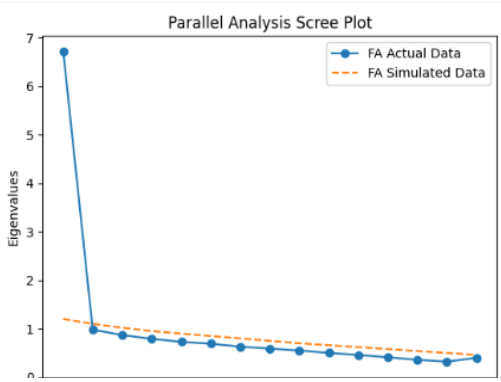

Figure 2 Scree plot

The scree plot shows that only the first factor has an eigenvalue significantly larger than those of the other factors and lies above the simulated eigenvalue line (parallel analysis). Meanwhile, the second factor and subsequent factors lie below the simulated line. These findings indicate that only one factor is worth retaining, so the 15-item instrument can be considered unidimensional.

Model fit test

Model fit tests were conducted to ensure that the 2PL model used was consistent with the participants' response data. These tests aim to verify that the applied model accurately describes the participants' response patterns. There are two approaches to conducting model fit tests: 1) the statistical approach and 2) the graphical approach. The results of the suitability index are listed in Table 4.

Table 4 Fit index

Model	Log-Likelihood	M2	df	RMSEA	SRMSR	TLI	CFI
Model 1 PL	-1284.527	436.812	90	0.072	0.148	0.941	0.944
Model 2 PL	-1196.304*	401.275	75	0.061*	0.082*	0.963*	0.968*

The fit index was assessed using the log-likelihood value, M2 statistic, RMSEA, SRMSR, TLI, and CFI. The analysis results showed that the 2PL model had a log-

likelihood value of −1196.304, which was higher (less negative) than the 1PL model's value of −1284.527. This indicates that the 2PL model explains the data better than the 1PL model. Additionally, the M2 statistic value for the 2PL model (401.275) is also smaller than that of the 1PL model (436.812), indicating improved model fit to the data. The RMSEA value for the 2PL model, at 0.061, falls into the "good" category and is lower than that of the 1PL model (0.072). The SRMSR value for the 2PL model (0.082) is also smaller than that of the 1PL model (0.148), indicating lower residual correlation. Furthermore, the incremental fit indices such as TLI and CFI for the 2PL model (TLI = 0.963; CFI = 0.968) are higher than those for the 1PL model (TLI = 0.941; CFI = 0.944), indicating a more optimal model fit.

The next statistical test is a relative statistical test. This test focuses on examining the values of the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The model with the lowest AIC and BIC values is considered the best-fitting model. The AIC, BIC, and Likelihood ratio test was shown in Table 5.

Table 5 AIC, BIC, Likelihood Ratio Test (LRT)

Model	Log-Likelihood	AIC	BIC	X^2	df	p
Model 1 PL	-1284.527	2599.054	2698.371			
Model 2 PL	-1196.304*	2432.608*	2581.582*	176.446	15	0.000*

Based on the AIC and BIC criteria in *Table 5 AIC, BIC, Likelihood Ratio Test (LRT)*, the 2PL model has an Akaike Information Criterion (AIC) value of 2432.608, which is lower than the 1PL model's value of 2599.054. A similar pattern is also evident in the Bayesian Information Criterion (BIC) values, where the 2PL model yields a value of 2581.582, which is smaller than the 1PL model's value of 2698.371. The results of the Likelihood Ratio Test (LRT) show a χ^2 value of 176.446 with 15 degrees of freedom and a significance level of $p = 0.000$ ($p < 0.05$).

Difficulty Level (b) and Distinctiveness (a)

The difficulty level of the items (b) in this study was categorized based on the guidelines proposed by Hambleton (Hambleton & Swaminathan, 1985), while the item discrimination (a), obtained from estimates using the 2PL IRT model, was interpreted according to the criteria proposed by Baker (Baker & Kim, 2004). The results of item difficulty (a) and discrimination (b) are listed in Table 6.

Table 6 Item difficulty (a) and discrimination (b)

Item	Index Parameter b	Interpretation	Index Parameter a	Interpretation
1	-0.78	Medium	1.28	Medium
2	-0.64	Medium	1.44	High
3	-0.52	Medium	1.31	Medium
4	-1.35	Easy	0.74	Medium
5	-1.18	Easy	0.82	Medium
6	1.24	Hard	0.69	Medium
7	-0.31	Medium	1.39	High
8	0.12	Medium	1.56	High
9	0.34	Medium	1.48	High
10	0.56	Medium	1.34	Medium
11	0.88	Medium	1.26	Medium
12	-0.15	Medium	1.52	High
13	0.67	Medium	1.29	Medium
14	1.42	Hard	0.67	Medium
15	1.18	Hard	0.73	Medium

Based on the results of Table 6, the 15 items ranged from −1.35 to 1.42. This range indicates that the difficulty levels of the items fall into the easy-to-difficult categories, with a predominance in the moderate category. Based on the interpretation criteria used, 10 test items (66.67%) fell into the moderate category, 2 test items (13.33%) fell into the easy category, and 3 test items (20.00%) fell into the difficult category. All test items fall into the moderate to high categories according to Baker's (2001) criteria. There are five items in the high category (items 2, 7, 8, 9, and 12), indicating excellent discriminatory power in distinguishing students based on their latent ability. Meanwhile, the other ten items fall into the moderate category, which still reflects adequate discriminatory power.

In the Two-Parameter Logistic (2PL) model, the item characteristic curve (ICC) is S-shaped (sigmoid), depicting the relationship between a test-taker's ability (θ) and the probability of answering correctly. Parameter (a) determines the slope of the curve; the larger the value of a, the steeper the curve. Parameter (b) indicates the point on the ability scale where the probability of answering correctly is 0.5. The value of b shifts the position of the curve to the left or right along the θ axis. The results of the item characteristic curve (ICC) are shown in Figure 3.

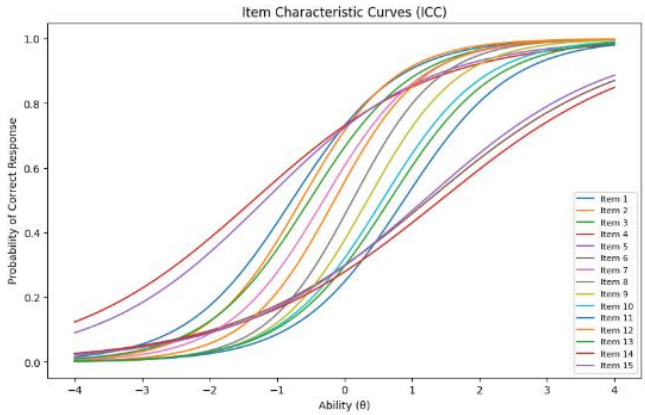

Figure 3 Item characteristic curves (ICC)

Based on Figure 3, each item exhibits a sigmoid logistic pattern that describes the relationship between student ability (θ) and the probability of answering correctly. Easy items (e.g., 4 and 5) shift to the left so that they can be answered by low-ability students, while difficult items (6, 14, 15) shift to the right and are more suitable for high-ability students. Most items are of moderate difficulty, centered around average ability. Items with high discriminative power (2, 7, 8, 9, 12) have steeper curves, making them more effective at distinguishing student ability compared to items with moderate discriminative power, although all are still considered adequate according to Frank B. Baker's criteria.

Based on the results of the content validity analysis using Aiken's V index, all 15 items in the Grade 12 high school physics formative test in Bulukumba Regency were found to meet the content validity criteria in the "very good" category (score > 0.75), with Aiken's V index values ranging from 0.80 to 0.90 and an average of 0.85. These findings indicate that these test items are appropriate for interpreting scores, ensure the representation of the learning objectives being measured, support fairness in assessment, and result in more accurate educational decision-making based on measurement data (Popham, 2017). The quality of achievement tests is highly determined by the developers' competence in designing high-quality items. To achieve this quality, each item must undergo a planned and systematic review process. This process aims to ensure that the construct being measured has been operationally defined and that test specifications are rationally

formulated so as to be consistent with the construct's definition (Haladyna, 2004). Thus, the review process provides a strong evidential foundation to support the validity of the test's use, as well as the interpretation of scores and participants' responses to the items.

High reliability is a crucial aspect because instruments with low reliability may produce inconsistent scores, causing students' abilities to appear to have changed when in fact they have not (Alim et al., 2025). This situation can lead to reduced assessment accuracy, difficulties in identifying learning barriers, and suboptimal determination of learning interventions (Ningrum, 2014). In the long term, the unreliability of an instrument can also have a negative impact on students' motivation and self-confidence because they may feel that the assessment they receive does not fairly reflect their abilities (Gompi et al., 2022). Therefore, a high reliability score for this instrument serves as an important indicator that the test possesses strong measurement quality and is suitable for assessing students' learning achievements.

The results of the Kaiser–Meyer–Olkin (KMO) test showed a value of 0.842, which falls within the "good" category; therefore, the sample size was deemed adequate for factor analysis. Furthermore, the results of Bartlett's Test of Sphericity showed a chi-square value of 1287.456 with 105 degrees of freedom and a significance level of 0.000 ($p < 0.05$). Thus, based on the results of these tests, it can be concluded that the research data meet the requirements for further analysis using the parallel analysis method.

The assumption of unidimensionality in IRT analysis requires that students' responses to each item be primarily influenced by a single latent factor, namely problem-solving in physics related to dynamic electricity. These findings indicate that the instrument with 15 items is unidimensional, meaning that all items measure the same construct. These results align with the concept of essential unidimensionality (Embretson & Reise, 2000; Reckase, 2009), which states that an instrument can be considered unidimensional if a single primary factor explains the largest proportion of item variance.

In testing for local independence, an instrument is said to meet the criterion of unidimensional if participants' responses to one item are independent of their responses to other items. The fulfillment of the unidimensionality assumption can serve as evidence that local independence has been achieved (Demars, 2010; Retnawati, 2014). Thus, the analysis results indicate that both assumptions, unidimensional and local independence, have been met in the data of this study, allowing student performance to be understood as the result of the influence of a single common construct.

Model fit was assessed using the log-likelihood value, M2 statistic, RMSEA, SRMSR, TLI, and CF (Kreitchmann et al., 2023). The analysis results showed that the 2PL model had a log-likelihood value of −1196.304, which was higher (less negative) than the 1PL model's value of −1284.527. This indicates that the 2PL model explains the data better than the 1PL model. Additionally, the M2 statistic for the 2PL model (401.275) is smaller than that of the 1PL model (436.812), indicating improved model fit to the data. In terms of goodness-of-fit indices, the 2PL model demonstrates better values than the 1PL model. The RMSEA value for the 2PL model is 0.061, which falls into the "good" category and is lower than that of the 1PL model (0.072). The SRMSR value for the 2PL model (0.082) is also smaller than that of the 1PL model (0.148), indicating lower residual correlation.

Based on the AIC and BIC criteria, the 2PL model has an Akaike Information Criterion (AIC) value of 2432.608, which is lower than the 1PL model's value of 2599.054. A similar pattern is also evident in the Bayesian Information Criterion (BIC) values, where the 2PL model yields a value of 2581.582, which is smaller than the 1PL model's value of 2698.371. These lower AIC and BIC values indicate that the 2PL model provides a more optimal balance between the level of fit to the data and the number of parameters used. The results of the Likelihood Ratio Test (LRT) showed a χ^2 value of 176.446 with 15 degrees of

freedom and a significance level of $p = 0.000$ ($p < 0.05$). These results indicate that there is a significant difference in fit between the 1PL model and the 2PL model, meaning that the 2PL model statistically provides a significant improvement in fit compared to the 1PL model. Based on these overall results, it can be concluded that the 2PL model is a more suitable and recommended model for use in data analysis in this study compared to the 1PL model.

The results of the item difficulty analysis revealed that most items fell into the moderate category, with difficulty parameters (b) ranging from −1.35 to 1.42. Specifically, there were two items classified as easy, several items in the moderate category, and three items in the difficult category. The difficulty level distribution, dominated by the moderate category, indicates that the developed test has sufficient ability (Rahman et al., 2021; Saepuzaman et al., 2022) to distinguish students' abilities across diverse proficiency levels. Furthermore, the presence of both easy and difficult items indicates that the instrument covers a range of ability levels, thereby potentially providing more comprehensive measurement information. Practically, an instrument with a varied difficulty distribution has the potential to support instructional decision-making (Mandinach & Schildkamp, 2021), such as identifying remedial needs for students with low ability and enrichment for students with high ability.

Teachers can use information regarding item difficulty levels and student performance to determine the type of feedback that best suits the quality of the work and the characteristics of the students. The results of this analysis can also serve as a basis for developing questions that probe students' understanding, identify their level of conceptual mastery, and function as a means of observing their comprehension. This information can further serve as a consideration for teachers in making instructional decisions, such as proceeding with the material, repeating explanations, or temporarily pausing instruction to clarify concepts. Thus, formative assessment analysis serves not only as a tool for evaluating learning outcomes but also as a diagnostic instrument that supports continuous and responsive instructional decision-making tailored to students' needs.

Within the framework of Item Response Theory (IRT), discriminant power represents an item's sensitivity to differences in participants' latent ability, while difficulty level indicates the item's position or location on the ability continuum. The combination of these two parameters allows researchers to identify items that are not only classified as easy or difficult but also effective or ineffective in accurately measuring ability (Hidayatullah et al., 2022). Thus, discrimination becomes an important indicator of item quality because a high discrimination value indicates that the item is capable of providing accurate information regarding differences in participants' abilities. Conversely, items with low discriminant validity tend to yield limited information and are therefore less optimal for use in measuring ability (Saepuzaman et al., 2021).

The results of the discrimination index (a) parameter estimate in the 2PL IRT model indicate that the quality of the developed test items is good, as all items fall into the moderate to high categories based on the Baker (2001). This suggests that each item has sufficient ability to distinguish between students with different levels of latent ability, thereby enabling the instrument to function effectively as a measurement tool. The presence of five items in the high category indicates that some items have very strong sensitivity in detecting differences in student ability.

When examined in terms of the difficulty parameter (b), all items in the high category fall within the moderate difficulty range. This finding aligns with measurement theory, which states that items with moderate difficulty tend to have more optimal discriminative power because they can identify ability differences across a broader range of students (Baker & Kim, 2004; Ebel & Frisbie, 1991; Lord, 1980). The high discriminative power of

these items may also be influenced by the alignment of the indicators with the competencies being measured, the clarity of the item wording, and the effective functioning of the distractors. Therefore, high-category items can be considered highly informative and worthy of retention in both the instrument and the item bank.

The remaining ten items in the moderate category still demonstrate adequate discriminative power and remain suitable for use in educational assessment contexts. In terms of difficulty, items in the moderate category exhibit a broader range of difficulty levels, namely easy, moderate, and difficult. Items with extreme difficulty levels (too easy or too difficult) tend to have lower discriminative power compared to moderate-difficulty items, as they fail to distinguish students optimally across the full range of ability. This is evident in items 4, 5, 6, 14, and 15, which have easy or difficult difficulty levels and relatively lower discriminative power compared to other items, items in the easy category (4 and 5), items in the difficult category (6, 14, and 15), and items in the moderate category (1, 2, 3, 7, 8, 9, 10, 11, 12, and 13). Nevertheless, the moderate category still indicates good item quality and remains suitable for use in educational assessment instruments. Minor improvements, such as enhancing the quality of distractors, clarifying the wording of the questions, or adjusting the difficulty level, can be made to optimize the discriminative power.

Unlike previous studies, which generally focused on testing the validity and reliability of instruments using the Classical Test Theory approach (Elviana, 2020; Fitraynsyah et al., 2024; Miftahir et al., 2025; Syahriani, 2021) or simply assessing the overall quality of the instruments, this study provides more in-depth information through a comprehensive psychometric analysis. The research findings indicate that, in addition to the instrument demonstrating good content validity (Baharuddin et al., 2025) and reliability (Nazharita et al., 2022), the analysis of item characteristics was also able to identify the level of difficulty and discriminant validity. Furthermore, model fit tests were conducted on the obtained data. The results show variations in difficulty levels and discriminant power, including model fit tests that align with the data (Reise & Moore, 2023; Wilson, 2005). These findings underscore the importance of rigorously testing the quality of instruments to make a greater contribution to the development of assessment and the improvement of the learning process.

CONCLUSION

Based on the results of an analysis using the 2PL Item Response Theory (IRT) model of the formative assessment instrument for physics problem-solving on the topic of dynamic electricity at high schools in Bulukumba Regency, it can be concluded that the test has good content validity, with all items showing an Aiken's V index above 0.75. The reliability of the instrument falls into the "good" category, indicating that the test has strong internal consistency.

The 2PL IRT analysis indicates that the difficulty levels of the 15 test items range from easy to difficult, with the majority falling into the moderate category (10 items), while a small portion are classified as easy (2 items) and difficult (3 items). This distribution is considered good because the majority of items are at the moderate level and there are no extreme items, meaning the instrument meets the expected difficulty characteristics in IRT-based measurement. Regarding the item discrimination index (a) of the 2PL IRT model, all items fall into the moderate to high categories, with five items in the high category and ten in the moderate category. There are no items in the low category, so the instrument has good discriminant power and is effective in distinguishing students' abilities.

The results of this study support the use of IRT 2PL model in the development of formative assessment instruments, particularly for measuring physics problem-solving

skills. These findings support the importance of item parameter analysis (a and b) in ensuring construct validity and the precision of measuring students' latent abilities. Thus, a systematic instrument development procedure (expert validation, empirical testing, and 2PL IRT analysis) is worthy of being adopted as a model for test development in other subjects.

The researchers acknowledge that this study has limitations, such as the instruments developed being limited to a single topic, electricity and dynamics. Another limitation is the focus on the use of the 2PL IRT model. Therefore, future item analysis research could incorporate the 3PL model and include discriminant parameters.

The implication of this study is the need for wider application of IRT-based item analysis in research on the development of educational instruments at the high school level. The use of the 2PL model in this study demonstrates that the quality of an instrument can be comprehensively evaluated through the parameters of difficulty and discriminant validity.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest.

AUTHOR CONTRIBUTIONS STATEMENT

All authors actively contributed to every stage of this study. CP was responsible for developing the initial concept, data collection, data analysis, and drafting the manuscript. KA and SI contributed to data validation, visualizing the research findings, and editing the final manuscript. All authors have read and approved the final version of this article.

AI DISCLOSURE STATEMENT

During the manuscript preparation process, the authors used ChatGPT to refine the language and carefully translated the text using DeepL. All aspects of the research design, data collection, data analysis, and drawing of conclusions were carried out entirely by the authors without the assistance of artificial intelligence (AI).

REFERENCES

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings, educational and psychological measurument. *Educational and Psychological Measurement, 45*(1), 131–142.

Alim, U. N., Syahrul, S., de Barros Cardoso, L. M. O., Ishak, S., & Asrifan, A. (2025). Analisis kualitas tes bahasa arab di indonesia: Studi systematic literature review tentang validitas, reliabilitas, tingkat kesukaran, dan daya beda. *Pepatudzu: Media Pendidikan dan Sosial Kemasyarakatan, 21*(2), 183-202.

Andrade, H. L., & Heritage, M. (2018). *Using formative assessment to enhance learning , achievement , and academic self-regulation.* ROutledge.

Andrade, H. L., Heritage, M., Cizek, G. J., Dick, W., Carey, L., Carey, J. O., Wilson, M., Madaus, G. F., College, B., Hill, C., Stufflebeam, D. L., Michigan, W., Finch, W. H., French, B. F., Ebel, R. L., Frisbie, D. A., Suen, H. K., Mehrens, W. A., Lehmann, I. J., … Denison, D. B. (1988). *Classroom assessment techniques a handbook for college teachers.pdf (N. J. Salkind, Ed.; 1st & 2nd ed., p. 1120)*. Jossey-.

Ankomah, F., & Nugba, R. M. (2020). Validation of test construction skills inventory through the lens of item response theory (IRT). *American Journal of Creative Education, 3*(2), 86-100.

Azwar, S. (2016). *Reliabilitas dan validitas (IV).* Pustaka Pelajar.

Arikunto, S. (2019). *Dasar-dasar evaluasi pendidikan (3rd ed.).* Bumi Aksara.

Baharuddin, H., Riskawati, R., & Fitrawardani, F. (2025). Identifikasi kemampuan prosedural dan konseptual siswa dalam memecahkan masalah matematika dengan model dina. *Jurnal Studi Guru dan Pembelajaran*, *8*(3), 2273-2282.

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

Baker, F. B., & Kim, S.-H. (2004). *Item response theory parameter estimation techniques (second).* Taylor & Francis Group. https://doi.org/10.4324/9781315391342-12

Beck, K. (2020). Ensuring content validity of psychological and educational tests – The role of experts. *Frontline Learning Research*, *8*(6), 1–37. https://doi.org/10.14786/flr.v8i6.517

Demars, C. (2010). *Item respon theory; Understanding statistics measurement (N. Beretvas, Ed.).* Oxford university Press.

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

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

Elviana, E. (2020). Analisis butir soal evaluasi pembelajaran pendidikan agama islam menggunakan program. *Jurnal Mudarrisuna, 10*(2), 58–74.

Fitraynsyah, M. A., Hilmiyati, F., & Sultan Maulana Hasanuddin, U. (2024). Peran tingkat kesukaran dan daya pembeda dalam analisis butir tes: Kajian literatur untuk pendidikan menengah. *Jrep :Jurnal Riset Dan Evaluasi Pendidikan*, *1*(4), 252–262.

Gompi, M., Bito, N., & Isa, D. R. (2022). Pengaruh model pembelajaran student facilitator and explaining terhadap hasil belajar sisiwa pada materi bentuk aljabar. *Jurnal Cendekia : Jurnal Pendidikan Matematika, 6*(3), 3287–3295. https://doi.org/10.31004/cendekia.v6i3.1739

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items (3rd ed.).* Mahwah, NJ: Lawrence Erlbaum Associates.

Hidayatullah, R., Safitri, M., & Suyanto, H. (2022). Analisis butir soal Bahasa Arab di MAN 2 Kota Padang. *Jurnal Evaluasi Pembelajaran Islam, 13*(1), 77–89.

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

Kreitchmann, R. S., de la Torre, J., Sorrel, M. A., Nájera, P., & Abad, F. J. (2023). Improving reliability estimation in cognitive diagnosis modeling. *Behavior Research Methods, 55*(7), 3446–3460. https://doi.org/10.3758/s13428-022-01967-5

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328–329

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

Magnus, B. (2018). The Basics of Item Response Theory Using R. *Measurement: Interdisciplinary Research and Perspectives*, *16*(3). https://doi.org/10.1080/15366367.2018.1462078

Mandinach, E. B., & Schildkamp, K. (2021). Misconceptions about Data-Based Decision Making in Education: An Exploration of the Literature. *Studies in Educational Evaluation, 69*, 100842

Miftahir, R., Andrina, M., Lestari, P. P., Indriyani, K. F., & Hidayanti, U. (2025). Analisis butir soal tes kemampuan siswa pada mata pelajaran manajemen perkantoran. *DIAJAR: Jurnal Pendidikan Dan Pembelajaran*, *4*(3), 337–346. https://doi.org/10.54259/diajar.v4i3.4029

Nazharita, E., Hidayat, R., & Rahmatudin, J. (2022). Domain Geometri Siswa Smp Menggunakan Rasch. *Pedagogy*, *10*(3), 989–1000.

Ningrum, D. O. S. (2014). Teknik guru dalam penyusunan soal, validitas, reliabilitas, kendala dan upaya dalam penyelesaian masalah penyusunan instrumen evaluasi

pembelajaran ipa biologi smp se-kabupaten probolinggo. Skripsi Universitas Jember. https://repository.unej.ac.id/handle/123456789/19137

Nitko, A. J., & Brookhart, S. M. (2011). *Educational Assessment of Students (P. A. Smith, Ed.; 6th ed.)*. Pearson.

Popham, W. J. (2017). *Classroom assessment: What teachers need to know (8th ed.)*. Pearson. www.pearsoned.com/permissions/.

Rahman, A., Rusnayati, H., & Muslim, M. (2021). Analysis of the characteristics of higher order thinking skills (HOTS) test on momentum and impulse for senior high school student using item response theory. *Jurnal Penelitian Fisika Dan Aplikasinya (JPFA), 11*(2), 127-137

Reckase, M. D. (2009). *Multidimensional item response theory (statistics for social and behavioral sciences).* Springer. https://doi.org/10.1007/978-0-387-89976-3

Reise, S. P., & Moore, T. M. (2023). *Item response theory*. In APA handbook of research methods in psychology: Foundations, planning, measures, and psychometrics (2nd ed., Vol. 1, pp. 809–835). American Psychological Association. https://doi.org/https://doi.org/10.1037/0000318-037

Retnawati, H. (2014). *Teori respon butir dan penerapanya*. Yogyakarta: Nuha Medika.

Saepuzaman, D., Istiyono, E., & Haryanto. (2022). Characteristics of fundamental physics higher-order thinking skills test using item response theory analysis. *Pegem Journal of Education and Instruction, 12*(4), 269-279. https://doi.org/10.47750/pegegog.12.04.28

Salkind, N. J., Frey, B. B., & Winter, R. J. (2020). *Statistic for people who (think they) hate statistic (7th ed.)*. SAGE.

Syahriani, S. (2021). Analisis butir soal ujian sekolah mata pelajaran bahasa indonesia kelas vi sdn kota bengkulu. *Jurnal Pembelajaran Dan Pengajaran Pendidikan Dasar, 5*(1), 198–209. https://doi.org/10.33369/dikdas.v5i1.16221

Syafii, A., Haryanto, Ahmad, I. F., & Fauziah, A. (2021). Analysis of items with item response theory (IRT) approach on final assessment for Al-Quran Hadith subjects. *Jurnal Pendidikan Agama Islam, 18*(1), 167–194. https://doi.org/10.14421/jpai.2021.181-09

Wilson, M. (2005). *Constructing measure: An item respon modeling approach*. Larence Erbaum Associates.

AUTHOR PROFILES

Candra Purnama
candra.purnama05@gmail.com.,
expertise in physics learning
Affiliation: Universitas Negeri Makassar

Kaharuddin Arafah
kahar.arafah@unm.ac.id,
expertise in measuring and evaluating physics learning
Affiliation: Universitas Negeri Makassar
Orcid

Suryadi Ishak
suryadi.ishak@unm.ac.id
expertise in measuring and evaluating mathematics learning
Orcid Shinta Scopus Google Schoolar