



Analysis of Validity, Reliability, Difficulty Level, and Discrimination Power of Students' Critical Thinking Skill Test on Phase F Fluid Materials in Senior High School

Nur Tasya Putri^{1*}, Emiliannur¹

¹Department of Physics, Universitas Negeri Padang, Jl. Prof. Dr. Hamka Air Tawar Padang 25131, Indonesia
Corresponding author. Email: Nurtasyaputri30@gmail.com

ABSTRACT

A high-quality assessment instrument is one that provides data that reflects actual conditions, is consistent and reliable, demonstrates varying levels of difficulty, and has a good discrimination index. Therefore, this study aims to test the validity, reliability, difficulty level, and analytical power of the items used to assess students' critical thinking skills. This study is descriptive and evaluative. The subjects were 31 11th-grade students at SMA Negeri 2 Painan. The data collection method used was a test. The study results indicated that for Grade XI, 17 items were deemed valid, whereas 1 item was invalid. The reliability coefficient for Grade XI was 0.877, demonstrating that the test instrument is reliable. Regarding difficulty level, all 18 items were classified as moderate. In terms of discrimination power, 5 items fell into the poor category, while 13 items were categorized as fair. Based on the analysis of physics test items on the topic of fluids for the 2024/2025 academic year at SMA Negeri 2 Painan, it was concluded that the 11th-grade test items were valid and reliable. This indicates that the test items can be used as an evaluation tool. Researchers recommend that teachers conduct trials on test items before giving them to students, in order to ensure the quality of the questions being tested.

Keywords: Critical Thinking Skills, Validity, Reliability, Difficulty Level, and Discrimination Power



Physics Learning and Education is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

I. INTRODUCTION

Twenty-first-century education aims to foster individuals' knowledge, skills, and attitudes through learning experiences, with an emphasis on mastering competencies needed to face future demands, also known as 21st-century skills [1]. There are three components of 21st-century skills needed by students, one of which is learning and innovation skills [2]. These skills help students adapt to an ever-changing environment, enabling them to become successful individuals in the future. These include core learning and innovation skills. These are considered essential skills to develop in students, as they encompass the capacity for critical thinking and responsible decision-making [3].

Critical thinking refers to a rational and reflective mode of thinking that centers on beliefs and actions. Decisions to be made. It is a process in which all knowledge and skills are utilized to solve emerging problems, make decisions, analyze assumptions, and conduct investigations or research based on the data and information obtained, resulting in the desired conclusions or information [4]. Critical thinking skills encompass six core subskills: interpretation, analysis, evaluation, inference, explanation, and self-regulation [5]. Modern education must foster these skills in students to prepare them for challenges and demands in daily life. Such skills can be cultivated and practiced through classroom learning activities.

One of the challenges faced by secondary school physics teachers is the low level of students' critical thinking skills, which are commonly assessed using essay tests. Tests are essential instruments in educational evaluation, as they provide objective information about students' understanding, competency mastery, and ability to apply knowledge in specific contexts [6]. Test results also serve as feedback for teachers to refine instructional strategies and for students to recognize their strengths and weaknesses [7].

However, test results are meaningful only when the assessment instrument meets established quality criteria, namely validity, reliability, difficulty level, and discrimination power. Item analysis is therefore necessary to ensure that essay questions accurately measure the intended competencies [8]. These four aspects constitute the quantitative evaluation of test items and are fundamental to improving the quality of assessment instruments and, consequently, educational outcomes.

Validity refers to the extent to which a test accurately measures the construct it is intended to assess [9], [10], [11]. Reliability, on the other hand, indicates the consistency and stability of measurement results when the assessment is administered repeatedly [12], [13], [14]. The relationship between validity and reliability can be explained as follows: validity is essential, while reliability is necessary because reliability supports validity [15].

In addition, high-quality test items should have an appropriate level of difficulty neither too easy nor too difficult. Difficulty level reflects the proportion of students who answer an item correctly and should be determined based on actual student performance rather than teacher assumptions [16]. Furthermore, a test item must possess adequate discrimination power, meaning it can effectively differentiate between high- and low-achieving students [17], [18].

Essay tests remain widely used because they are effective in assessing higher-order thinking skills, such as analysis, synthesis, evaluation, and logical argumentation. To ensure the quality of such assessments, teachers must conduct systematic item analysis to identify good, fair, and poor questions, as well as to detect weaknesses requiring revision [19], [20]. Therefore, conducting a thorough item analysis is crucial to determining the overall quality of the test.

II. METHOD

This descriptive study aimed to test the validity, reliability, difficulty level, and discrimination power of test items intended to assess students' critical thinking skills. To achieve this goal, the designed test items were administered to 31 eleventh-grade students of SMA Negeri 2 Painan. The test items analyzed in this study were critical thinking essay questions, with test guidelines aligned with the critical thinking stages based on the facione indicators [5]. The test items measured high school students' critical thinking skills related to the topic of Fluids, which consists of three subtopics: Pascal's Law, Archimedes' Principle, and Bernoulli's Principle, with a total of 18 test items. The validity of each test item was evaluated using the Pearson Product Moment correlation, as presented below.

$$r_{xy} = \frac{n \sum XY - (\sum X)(\sum Y)}{\sqrt{[n \sum X^2 - (\sum X)^2][n \sum Y^2 - (\sum Y)^2]}} \quad (1)$$

[21]

Information:

r_{xy} = Correlation coefficient between variables X and Y

X = variable group value X

Y = variable group value Y

n = number of participants

The validity criteria are shown in Table 1.

Table 1. Validity Test Criteria

Criteria	Description
$r_{calculated} > r_{table}$	Valid
$r_{calculated} \leq r_{table}$	Invalid

A test is considered reliable when multiple administrations yield relatively consistent results. To evaluate the reliability of the essay-based test, the Alpha formula was applied, as shown in the following equation.

$$r_{11} = \frac{n}{n-1} \left(1 - \frac{\sum S_f^2}{S_t^2} \right) \quad (2)$$

[21]

Information :

r_{11} = reliability coefficient

n = number of participants

$\sum S_t^2$ = Sum of the variances of each item

S_t^2 = total variance

The criterion is determined by comparing the r_{11} value with the product-moment correlation table. The r_{11} value is considered significant if $r_{11} > r_{table}$ at a 5% significance level. The criteria for the reliability coefficient are shown in Table 2.

Table 2. Criteria for Cronbach's Alpha Coefficient

No.	Reliability Coefficient	Interpretation
1.	$0,00 \leq r_i \leq 0,50$	Low
2.	$0,50 \leq r_i \leq 0,70$	Keep
3.	$0,70 \leq r_i \leq 0,90$	High
4.	$0,90 \leq r_i \leq 1,00$	Very high

[22]

The level of difficulty indicates whether a test item is classified as difficult, moderate, or easy to answer. An item's difficulty level can serve as a measure of its quality and indicate whether it requires revision. The difficulty index of a test item can be calculated using the following formula:

$$TK = \frac{\bar{X}}{SMI} \quad (3)$$

[23]

Description:

TK = Difficulty level

\bar{X} = Average score

SMI = Ideal maximum score

The difficulty level of a test item can be determined using the interpretation presented in Table 3.

Table 3. criteria for the level of difficulty of the questions

Difficulty Index	Category
$0,00 < IK \leq 0,30$	Difficult
$0,31 < IK \leq 0,70$	Keep
$0,71 < IK \leq 1,00$	Easy

[23]

A high-quality assessment instrument should be capable of distinguishing between students with high and low levels of ability in answering test items. The discrimination power of an item indicates its effectiveness in differentiating high-performing students from low-performing ones [24]. In other words, discrimination power reflects the degree to which a test item can separate students with strong abilities from those with weaker abilities [25]. The formula used for each test item is as follows:

$$DP = \frac{\bar{X}_A - \bar{X}_B}{SMI} \quad (4)$$

[23]

Information :

DP = Item discrimination index

\bar{X}_A = Average answer score of students in the upper

\bar{X}_B = Average answer score of students in the lower group

SMI = Ideal Maximum Score

The classification of item discrimination indices is presented in Table 4.

Table 4. Discrimination power index

Discrimination Index	Category
$0,70 < DP \leq 1,00$	Excellent
$0,40 < DP \leq 0,70$	Nice
$0,20 < DP \leq 0,40$	Enough
$0,00 < DP \leq 0,20$	Ugly
$DP \leq 0,00$	Very Ugly

[21]

III. RESULTS AND DISCUSSION

Results

The findings of the analysis of the critical thinking skills test items are presented as follows.

Results of the Test Item Validity Analysis

Before conducting the research, the test questions and their answer keys were first developed and validated by the supervisor. A total of 18 test items were validated and categorized as valid. The validity results of the test item trials as shown in Table 5 below.

Table 5. Results of Validity Calculation

Test Item Validity Analysis				
No.	r_{tabel}	r_{hitung}	Conclusion	Remark
1.	0,344	0,38	Valid	Used
2.	0,344	0,358	Valid	Used
3.	0,344	0,613	Valid	Used
4.	0,344	-0,044	Invalid	Revised
5.	0,344	0,607	Valid	Used
6.	0,344	0,652	Valid	Used
7.	0,344	0,529	Valid	Used
8.	0,344	0,677	Valid	Used
9.	0,344	0,680	Valid	Used
10.	0,344	0,542	Valid	Used
11.	0,344	0,455	Valid	Used
12.	0,344	0,614	Valid	Used
13.	0,344	0,733	Valid	Used
14.	0,344	0,811	Valid	Used
15.	0,344	0,754	Valid	Used
16.	0,344	0,777	Valid	Used
17.	0,344	0,760	Valid	Used
18.	0,344	0,724	Valid	Used

Based on the analysis of the 18 pretest and posttest items, out of the 18 questions that were tested, 17 items were found to be valid specifically items number 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, and 18 while one item, number 4, was deemed invalid. From the explanation above, the 17 valid items indicate that the majority of the questions successfully measured students' understanding of the fluids material. The valid test items demonstrate that these questions were able to provide an accurate depiction of students' comprehension levels in relation to the predetermined learning objectives.

Despite the overall results, the identification of one invalid item indicates the need for further improvement. A comprehensive analysis of the item is required to determine its alignment with validity criteria and to identify the underlying causes of its invalidity.

Results of Test Item Reliability Analysis

The researcher conducted a reliability test on the critical thinking skills questions, which consisted of 18

essay items for the Physics subject in Grade XI, with a total of 31 students. Instrument reliability was examined through the application of Cronbach's Alpha, with reliability established when the coefficient value is 0.65 or higher.

Table 6. Results of the Instrument

Reliability Statistics	
Cronbach's Alpha	Number of Question Items
0,877	18

Good reliability is characterized by a small margin of measurement error and consistent results when the test is administered multiple times. If the obtained reliability value falls between $0,90 \leq r_i \leq 1,000$ or $0,70 \leq r_i \leq 0,90$, the test items are considered reliable and will be used in the research instrument. If the reliability value is moderate $0,50 \leq r_i \leq 0,70$, the items will still be used but will need to be revised first. However, if the reliability value is low $0,00 \leq r_i \leq 0,50$, the items will not be used.

Results of Test Item Difficulty Level Analysis

The researcher conducted a test on critical thinking skills questions consisting of 18 essay items for the Physics subject in Grade XI, with a total of 31 students.

Table 7. Results of Item Difficulty Level Calculation (Trial Test)

Item No	Difficulty Coefficient	Category
1	0,43	Keep
2	0,69	Keep
3	0,64	Keep
4	0,80	Keep
5	0,65	Keep
6	0,66	Keep
7	0,59	Keep
8	0,65	Keep
9	0,70	Keep
10	0,64	Keep
11	0,64	Keep
12	0,56	Keep
13	0,47	Keep
14	0,52	Keep
15	0,47	Keep
16	0,46	Keep
17	0,43	Keep
18	0,62	Keep

Based on Table 3, the questions used in the study are those with a difficulty index ranging from $0,31 < IK \leq 0,70$. From Table 7, it can be observed that the difficulty levels of all 18 test items falls within the moderate category. Therefore, all 18 questions are considered appropriate and feasible to be used in this research.

Results of Item Discrimination Power Analysis

The analytical power of the questions can be seen in Table 8.

Table 8. Results of the Item Discrimination Power Analysis

No.	Discrimination Index	Category
1	0,21	Fair
2	0,14	low
3	0,21	Fair
4	0,05	Low
5	0,25	Fair
6	0,21	Fair
7	0,21	Fair
8	0,09	Low
9	0,28	Fair

No.	Discrimination Index	Category
10	0,05	Low
11	0,12	Low
12	0,21	Fair
13	0,30	Fair
14	0,34	Fair
15	0,27	Fair
16	0,27	Fair
17	0,31	Fair
18	0,28	Fair

Based on Table 8, 13 items demonstrate fair discrimination power, while 5 items are classified as having poor discrimination power. The five items in the poor category were subsequently revised.

Table 9. Item Analysis Results

No	Validity	Reliability Level	Discrimination Power	Difficulty Level	Classification	Further treatment
1	Valid	High	Fair	Moderate	Accepted	Accepted
2	Valid		Poor	Moderate	Rejected	Revised
3	Valid		Fair	Moderate	Accepted	Accepted
4	Invalid		Poor	Easy	Rejected	Revised
5	Valid		Fair	Moderate	Accepted	Accepted
6	Valid		Fair	Moderate	Accepted	Accepted
7	Valid		Fair	Moderate	Accepted	Accepted
8	Valid		Poor	Moderate	Rejected	Revised
9	Valid		Fair	Moderate	Accepted	Accepted
10	Valid		Poor	Moderate	Rejected	Revised
11	Valid		Poor	Moderate	Rejected	Revised
12	Valid		Fair	Moderate	Accepted	Accepted
13	Valid		Fair	Moderate	Accepted	Accepted
14	Valid		Fair	Moderate	Accepted	Accepted
15	Valid		Fair	Moderate	Accepted	Accepted
16	Valid		Fair	Moderate	Accepted	Accepted
17	Valid		Fair	Moderate	Accepted	Accepted
18	Valid		Fair	Moderate	Accepted	Accepted

Based on Table 9, out of the 18 test items that were tried out, 17 items were found to be valid. The reliability coefficient obtained, which is 0.877, is classified within the very high reliability category. The difficulty level analysis of the 18 items showed that all questions were in the moderate category. Furthermore, 13 items had a fair discrimination power, while 5 items were categorized as poor and were subsequently revised. It can be inferred that the test items are suitable and feasible for use in this study.

Discussion

This research aims to examine essay-type questions developed to assess students' critical thinking skills. The questions analyzed in this study consist of descriptive questions assessing critical thinking skills, developed according to guidelines aligned with the stages of critical thinking [5]. The following is the discussion of each characteristic of the test items:

1. Validity

Item validity was determined using the Pearson Product Moment correlation by comparing the calculated correlation coefficient with the r -table value at a 5% significance level. With 31 students participating in the trial ($N = 31$), the r -table value was 0.344. Items with $r_{\text{calculated}} > r_{\text{table}}$ were classified as valid, while those with $r_{\text{calculated}} \leq r_{\text{table}}$ were considered invalid.

The test instrument developed in this study consisted of open-ended essay questions designed to measure students' critical thinking skills. A test with high validity is essential, as it ensures that the instrument accurately measures the intended construct and provides reliable information about students' actual abilities [26]. Valid assessment results are crucial for teachers in making appropriate instructional decisions. Conversely, instruments with low validity may lead to misleading interpretations of students' learning outcomes and negatively affect both teaching and learning processes.

The analysis of the Critical Thinking Skills Test, which comprised 18 essay items, showed that 17 items were classified as valid, while 1 item was classified as invalid. These findings are consistent with previous research reporting that the majority of developed items met validity criteria [27]. The invalid item was likely caused by an imbalance in item difficulty, as items that are too easy or too difficult tend to yield low item-total correlation values when analyzed using the Pearson Product Moment formula [25].

Overall, the results indicate that the Critical Thinking Skills Test demonstrates good validity. Valid items can be retained and stored in the item bank, while invalid items should be revised or discarded before reuse.

2. Reliability

Reliability reflects the consistency of an assessment instrument when administered repeatedly [28]. Reliability is determined based on the Alpha value obtained from the calculation output. The reliability analysis resulted in a Cronbach's Alpha coefficient of 0.877, which is classified as having high reliability, because 0.877 falls within the interval $0,70 \leq r_i \leq 0,90$ [22]. This result is comparable to previous research reporting similarly high reliability coefficients for essay-based critical thinking tests [27].

The high alpha value indicates that the test items are internally consistent and capable of producing stable results across administrations. Therefore, the developed instrument can be considered reliable and suitable for measuring students' critical thinking skills [29], [30].

3. Level of Difficulty

Item difficulty is an important indicator of test quality, as overly easy or overly difficult items may fail to function optimally. The analysis revealed that all 18 items fell within the medium difficulty category, which is considered ideal for educational assessments. Items with moderate difficulty levels are more effective in engaging students and differentiating their abilities [31].

The difficulty index values were within the recommended range of $0,31 < IK \leq 0,70$, indicating that the test items are neither too simple nor too demanding. Consequently, the Critical Thinking Skills Test demonstrates good quality in terms of difficulty level and is appropriate for inclusion in an item bank for future use [32].

4. Discrimination Power

Discrimination power refers to an item's ability to distinguish between students with high and low levels of ability [6]. The analysis showed that 13 items were classified as having fair discrimination power, while 5 items were categorized as low. Similar patterns have been reported in previous studies, where some items exhibited limited discrimination despite being valid and moderately difficult [27].

The low discrimination power observed in several items may be attributed to conceptual and instructional factors. Some items may have been too straightforward or focused on surface-level understanding, allowing both high and low performing students to answer them correctly. In addition, limited variation in students' prior knowledge or instructional emphasis on similar problem types may have reduced the items' ability to differentiate levels of critical thinking performance.

Despite these findings, the overall discrimination power of the test can be considered satisfactory, as the majority of items were able to distinguish between students who had mastered the material and those who had not. Items with fair discrimination power can be retained and stored in the item bank, while items with poor discrimination power should be revised to increase cognitive demand, clarify prompts, or better align with higher order thinking indicators. Items with negative discrimination indices should be eliminated, as they do not function effectively as assessment tools [33].

IV. CONCLUSION

The quality of the items developed to assess students' critical thinking skills falls within the high validity category. The reliability coefficient of the test is classified as strong, indicating that the instrument consistently measures students' critical thinking skills. The difficulty level of the items is categorized as moderate, demonstrating the test's ability to effectively differentiate students' abilities. Furthermore, the results of the item

discrimination analysis indicate that most items exhibit good discriminatory power, enabling the instrument to distinguish between students who have mastered the concepts and those who have not fully understood the material. Overall, these findings indicate that the instrument is effective and suitable for long-term evaluation purposes.

REFERENCES

- [1] Khalifah, I., Sakti, I., & Sutarno, S. (2021). Pengembangan LKPD berbasis Project Based Learning untuk melatih keterampilan berpikir kritis pada materi induksi elektromagnetik. *DIKSAINS: Jurnal Ilmiah Pendidikan Sains*, 1(2), 69–80.
- [2] Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times*. Jossey-Bass.
- [3] Redhana, I. W. (2010). Pengaruh model pembelajaran berbasis peta argumen terhadap keterampilan berpikir kritis siswa pada topik laju reaksi. *Jurnal Pendidikan dan Pengajaran*, 43(1), 141–148.
- [4] Ariyana, Y., Bestary, R., Pudjiastuti, R., & Zamroni. (2018). *Buku pegangan pembelajaran berorientasi pada keterampilan berpikir tingkat tinggi*. Direktorat Jenderal Guru dan Tenaga Kependidikan, Kementerian Pendidikan dan Kebudayaan.
- [5] Facione, P. A. (2013). *Critical thinking: What it is and why it counts*. Measured Reasons and California Academic Press.
- [6] Arikunto, S. (2019). *Prosedur penelitian*. Rineka Cipta.
- [7] Indah, M., Karoma, K., & Rusdi, A. (2021). Analisis tes butir soal guru dalam mata pelajaran Pendidikan Agama Islam di SMP Negeri 8 Palembang. *Muaddib: Islamic Education Journal*, 4(1), 22–30.
- [8] Sugiyono. (2021). *Metode penelitian kuantitatif, kualitatif, dan R&D* (Edisi ke-2). Alfabeta.
- [9] Taherdoost, H. (2016). Validity and reliability of the research instrument: How to test the validation of a questionnaire/survey in a research. *International Journal of Academic Research in Management*, 5(3), 28–36.
- [10] Sullivan, G. M. (2011). A primer on the validity of assessment instruments. *Journal of Graduate Medical Education*, 3(2), 119–120.
- [11] Muzaffir, A. (2016). Validitas tes dan kualitas butir soal. *Lisanuna: Jurnal Bahasa Arab dan Pembelajaran*, 5(1).
- [12] Ghozali, I. (2018). *Aplikasi analisis multivariate dengan program IBM SPSS 25* (Edisi ke-9). Badan Penerbit Universitas Diponegoro.
- [13] Bajpai, S., & Bajpai, R. (2014). Goodness of measurement: Reliability and validity. *International Journal of Medical Science and Public Health*, 3(2), 112–116.
- [14] Surapranata, S. (2006). *Analisis, validitas, reliabilitas, dan interpretasi hasil tes*. Remaja Rosdakarya.
- [15] Sukiman. (2012). *Pengembangan sistem evaluasi*. Insan Madani.
- [16] Boopathiraj, C., & Chellamani, K. (2013). Analysis of test items on difficulty level and discrimination index in the test for research in education. *International Journal of Social Science & Interdisciplinary Research*, 2(2), 189–193.
- [17] Hendriana, H., & Sumarmo, U. (2014). *Penilaian pembelajaran matematika*. Refika Aditama.
- [18] Kocdar, S., Karadag, N., & Sahin, M. D. (2016). Analysis of the difficulty and discrimination indices of multiple-choice questions according to cognitive levels in an open and distance learning context. *TOJET: The Turkish Online Journal of Educational Technology*, 15(4).
- [19] Hayati, S., & Lailatussaadah, L. (2016). Validitas dan reliabilitas instrumen pengetahuan pembelajaran aktif, kreatif, dan menyenangkan (PAKEM) menggunakan model Rasch. *Jurnal Ilmiah Didaktika*, 16(2).
- [20] Farida, & Musyarofah, A. (2021). Validitas dan reliabilitas dalam analisis butir soal. *Al-Mu'arrib: Jurnal Pendidikan Bahasa Arab*, 1(1), 34–44.
- [21] Sundayana, H. R. (2020). *Statistika penelitian pendidikan*. Alfabeta.
- [22] Hinton, P. R., Brownlow, C., McMurray, I., & Cozens, B. (2004). *SPSS explained*. Routledge.
- [23] Lestari, K. E., & Yudhanegara, M. R. (2017). *Penelitian pendidikan matematika*. Refika Aditama.
- [24] Sundayana, R. (2016). *Statistika penelitian pendidikan*. Alfabeta.
- [25] Arikunto, S. 2010. *Prosedur Penelitian Suatu Pendekatan Praktik*. Jakarta: Rineka Cipta
- [26] Kuseiri, & Suprananto. (2012). *Pengukuran dan penilaian pendidikan*. Graha Ilmu.
- [27] Kamalia, N., & Wasis. (2021). Analisis profil keterampilan berpikir kritis peserta didik SMA dalam menyelesaikan soal fluida statis. *Jurnal Inovasi Pendidikan Fisika*, 10(1), 90–98.
- [28] Mahrawi, Usman, & Maulida, N. (2021). Pengembangan instrumen asesmen critical thinking skills pada materi sistem ekskresi. *Journal of Mathematics and Natural Science Education*, 2(2), 80–95.
- [29] Sugiyono, (2013). *Metode Penelitian Pendidikan: (Pendekatan Kuantitatif, Kualitatif, Dan R&D)*. Bandung: Alfabeta
- [30] *Tes implementasi kurikulum 2004*. (2004). Remaja Rosdakarya.
- [31] Arifin, Z. (2013). *Evaluasi pembelajaran*. Remaja Rosdakarya.
- [32] Susanto, H., Rinaldi, A., & Novalia. (2015). Analisis validitas, reliabilitas, tingkat kesukaran, dan daya beda pada butir soal ujian akhir. *Al-Jabar: Jurnal Pendidikan Matematika*, 6(2), 203–217.

- [33] Sudijono, A. (2011). *Pengantar evaluasi pendidikan*. RajaGrafindo Persada.