An Analysis of the Quality of the Summative Assessment Questions in Pancasila Education for the Second Semester of Sixth Grade Elementary Schools in Bengkulu City

Atika Susanti[1], Dwi Anggraini[2], Abdul Muktadir[3]

[1,2]Pendidikan Guru Sekolah Dasar, Universitas Bengkulu, Bengkulu, Indonesia
[3]Magister Pendidikan Dasar, Universitas Bengkulu, Bengkulu, Indonesia
*Corresponding author, email: atikasusanti@unib.ac.id

Keywords
Question Quality; Summative Assessment; Pancasila Education; Elementary School.

Abstract
This study aims to analyze the quality of the summative assessment questions for Pancasila Education in the second semester of grade VI at elementary schools in Bengkulu City. The method used is a mixed-methods approach with a concurrent embedded strategy, integrating both quantitative and qualitative approaches simultaneously. This study also employs a quadrant IV research design. The analysis was conducted qualitatively and quantitatively to determine the extent to which the questions accommodate students' cognitive levels as well as technical aspects such as difficulty level, discriminative power, distractor effectiveness, validity, and reliability. Qualitative analysis showed that the questions covered various cognitive levels from C1 to C6 with a varied proportion, supporting the development of students' critical thinking skills. Quantitative analysis revealed that most questions had an easy difficulty level of 62.5%, medium 32.5%, and hard 5%, with discriminative power mostly in the moderate to good category at 82.5%. However, distractor effectiveness was still low (72.5% ineffective) and question validity was not yet optimal (47.5% invalid). The reliability of the question bank was high, with a Cronbach's alpha of 0.785. Additionally, the dominance of questions in the LOTS domain at 52.5% indicates more emphasis on basic knowledge assessment, while HOTS questions comprised only 22.5%. This study recommends improving distractors and increasing the proportion of HOTS questions to optimize the assessment in fostering students' thinking skills in line with the demands of the 21st-century curriculum.

1. Introduction

Pancasila Education in elementary schools aims to shape students' character based on the noble values of Pancasila. According to Haduong et al. (2024), Pancasila Education helps develop civic knowledge, skills, and attitudes. Greany (2024) adds that students are not only expected to understand these values cognitively but also to internalize them in their daily lives. Civic attitudes and values are essential in shaping conscious and responsible citizens (Anderson, 2023). Therefore, effective civic education must balance attitudes, knowledge, and skills (Crittenden & Levine, 2018), thereby forming individuals who are prepared to actively and reflectively face the challenges of democracy.

Summative assessment serves as a crucial tool for measuring the achievement of these competencies. However, the quality of the assessment questions used in learning evaluation often fails to meet the established standards in terms of construct, content, language, validity, reliability, difficulty level, discrimination index, and the effectiveness of distractors. According to Crisp et al. (2019), poorly constructed test items can provide a misleading picture of students' abilities and may even hinder the overall teaching and learning process. Kwok & Kwan (2025) state that clear item instructions such as the use of explicit and commonly understood cue words greatly assist students in answering questions. Therefore, improving the quality of assessment items is essential so that the evaluation process can yield accurate results and support the optimal development of students' competencies.

According to Prijowuntato (2020), tests or evaluation instruments are tools that teachers can use to assess how well students have understood the material taught during the learning process. Widodo (2021) states that a test item can be considered good if it meets certain criteria or principles, enabling it to produce accurate data aligned with the intended evaluation objectives. Rifana et al. (2024) explain that test evaluation includes three main aspects: content validity, which examines the relevance of the material being tested; construct validity, which assesses the appropriateness of the item construction; and the reliability of the test itself. Arikunto (2021) also emphasizes that in evaluation activities, a quality test is one that accurately reflects the actual condition of the students.

According to Farida & Musyarofah (2021), in order to determine the quality of a test, it is necessary to conduct an item quality analysis. The purpose of this analysis is to classify test items based on their quality whether they are categorized as good, fairly good, or unsuitable for use (Hasibuan et al., 2024). In practice, item analysis needs to take into account several aspects such as validity, reliability, objectivity, ease of use, and efficiency (Magdalena, 2020; Sutikno, 2021). Damayanti et al. (2023) and Ramadhan et al. (2023) assert that a test item can be considered of high quality if it meets indicators such as item difficulty, discriminating power, and the effectiveness of answer choices (distractors). This reinforces the importance of conducting a comprehensive item evaluation, covering item difficulty, discrimination index, distractor quality, content and construct validity, and overall instrument reliability.

Several previous studies have evaluated learning outcomes and item analysis in Pancasila and Civic Education. Adawiah & Maulana (2024) found that the AnBuso version 8.0 program was effective in improving student learning outcomes, while Mbana et al. (2024) reported that the application had a usability score of 85, which falls into the "Excellent" category. Other research also highlights the need to revise questions that measure higher-order thinking skills (Maryani et al. 2021). In addition, assessment is considered essential to measure students' knowledge, attitudes, and skills (Harahap, 2024), and printed textbooks are necessary to reinforce students' understanding (Marsudi & Sunarso, 2019). Research by Kusumawati et al. (2025) also shows that, overall, the test items had good quality, with moderate difficulty level, good discriminating power, and effective distractors.

Nevertheless, several previous studies have indicated that the quality of assessment items in education remains a common issue. Ramadhan et al. (2024) revealed that Pancasila Education assessment items in some elementary schools have not fully considered the aspects of item discrimination and alignment with learning indicators. In Bengkulu City in particular, there has been limited research specifically analyzing the quality of summative assessment items for Pancasila Education based on the Merdeka Curriculum at the elementary school level. In fact, the evaluation of assessment instruments is crucial to ensure that the test items used are truly capable of measuring students' competencies objectively and fairly.

Research related to Pancasila education at the elementary school level has shown significant developments in various aspects of learning, especially within the context of the Merdeka Curriculum and differentiated learning approaches. Susanti et al. (2023) highlighted the implementation of a project to strengthen the Pancasila student profile, emphasizing character development through the Merdeka Curriculum at the elementary level, which was further supported by an analysis of strategies to enhance critical reasoning dimensions at SD Negeri 44 Kota Bengkulu (Susanti et al., 2023). Furthermore, research in 2024 by Susanti et al. examined students' learning style profiles to support differentiated instruction in the Merdeka Curriculum, alongside developing diagnostic assessments for Pancasila education for fourth-grade students at SDN 64 Bengkulu Tengah, demonstrating the need for valid and effective evaluation instruments (Susanti et al., 2024; 2025).

In addition, efforts to develop character-based learning media such as the Gurita media to improve cognitive learning outcomes on the practice of Pancasila principles also became a focus of study (Susanti & Muktadir, 2025). Another study by Susanti et al. (2024) on the analysis of differentiated learning approach implementation in fifth grade at SD Negeri 44 Kota Bengkulu underscored the importance of designing test instruments that not only measure basic knowledge but also critical thinking skills and student character. Based on this background, research on the analysis of Pancasila test items is highly relevant and contemporary as it plays a crucial role in ensuring the quality and effectiveness of learning evaluation instruments aligned with curriculum demands and the characteristics of today's students, thereby supporting the optimal achievement of Pancasila education goals.

The urgency of this study lies in the importance of providing feedback to teachers and schools regarding the quality of the assessment items used. This feedback can support the improvement and development of higher-quality test items in the future. Furthermore, this research can serve as a reference for efforts to improve the quality of education, particularly in strengthening student character through measurable and standardized Pancasila Education.

Based on the issues previously outlined, and in order to assist teachers in reviewing the appropriateness of the assessment items to be used, the researcher conducted a study on the quality analysis of summative assessment items for the second semester of sixth-grade Pancasila Education in elementary schools in Bengkulu City. This study employs both qualitative and quantitative approaches. Through this analysis, it is expected that high-quality test items can be identified items that are capable of optimally measuring student learning outcomes.

2. Method

The research employed a mixed methods approach, integrating both quantitative and qualitative methodologies. In this study, the qualitative approach served as the primary method, while the quantitative approach functioned as a secondary, supporting method. The research design followed the Quadrant IV model, with data collection techniques including measurement and document analysis.

The measurement technique was used to collect students' responses, whereas document analysis was applied to gather materials such as Pancasila Education test items for sixth grade, test blueprints, answer keys, and students' answer sheets. The data collection instrument consisted of a summative assessment test for the second semester of sixth grade, obtained from SD Negeri 44 Bengkulu City. The test comprised 40 multiple-choice items.

Data analysis techniques involved both qualitative and quantitative approaches. Qualitative analysis focused on the aspects of content, construction, and language, which were evaluated using a multiple-choice item assessment rubric. On the other hand, quantitative analysis included evaluating item difficulty level, discrimination index, distractor effectiveness, validity, and reliability. Additionally, the analysis examined students' levels of understanding in answering the items, based on Bloom's Taxonomy, covering the categories of knowledge (C1), comprehension (C2), application (C3), analysis (C4), synthesis (C5), and evaluation (C6).

Table 1. Difficulty Level Index

Range	Description
0,00 – 0,30	Difficult
0,31 – 0,70	Moderate
0,71 – 1,00	Easy

(Sudjana, 2009)

The Difficulty Level Index is a classification used to determine how challenging a test item is for students. Based on the index, items are categorized into three levels: difficult (0.00–0.30), moderate (0.31–0.70), and easy (0.71–1.00). This categorization helps educators evaluate whether the test items are appropriately balanced to assess a range of student abilities. Ideally, a good test includes a mix of difficulty levels to ensure fairness and comprehensiveness in measuring student understanding. An overrepresentation of items in one category, such as too many easy or difficult questions, can reduce the effectiveness of the test as an evaluation tool. Therefore, analyzing the difficulty index is essential in test development and refinement processes.

Table 2. Discrimination Index

Discrimination Index Item	Description
0 – 0,20	The item has a weak discrimination index
0.21 – 0.40	The item has a moderate discrimination index
0.41 – 0.70	The item has a good discrimination index
0.71 – 1.00	The item has a very strong discrimination index
Negative Value	The item has a very poor discrimination index

(Arikunto, 2003)

The discrimination index indicates the extent to which a test item can differentiate between students with high and low performance. According to standard classification, an item with a discrimination index below 0.20 is considered weak and may not effectively distinguish between varying levels of student ability. A value between 0.21 and 0.40 is considered moderate, while values between 0.41 and 0.70 indicate good discrimination. Items with values above 0.70 have very strong discrimination power. Conversely, a negative discrimination index suggests that more low-achieving students answered the item correctly than high-achieving students, making the item misleading and in need of revision.

3. Results and Discussion

3.1 Qualitative Analysis

The analysis of the test items indicates that the questions accommodate various levels of students' cognitive abilities, ranging from lower-order to higher-order thinking skills. There are 9 items in the C1 (Remembering) category questions 1, 6, 9, 11, 15, 17, 18, 20, and 21. These items assess students' ability to recall facts and basic information, such as definitions, dates, and key terms related to the history and values of Pancasila. Next, 12 items fall under the C2 (Understanding) category questions 2, 4, 5, 10, 19, 22, 26, 29, 30, 31, and 33. These questions test students' ability to comprehend concepts and relate Pancasila values to real-life attitudes and actions in everyday situations.

For the higher-order thinking level C3 (Applying), there are 10 items questions 3, 7, 8, 12, 14, 16, 23, 24, 28, and 38. In this category, students are required to apply their understanding of norms, values, and Pancasila principles in practical contexts such as school and community environments. There are 4 items in the C4 (Analyzing) category questions 13, 25, 32, and 34. These questions require students to analyze statements and distinguish between those that align or do not align with Pancasila values or prevailing norms, thus encouraging critical and in-depth thinking. In the C5 (Evaluating) category, there are 4 items questions 27, 35, 36, and 37. These items assess students' ability to judge and evaluate behaviors and actions based on Pancasila values and applicable rules, prompting reflective thinking and value-based assessment of social situations.

Finally, in the highest cognitive level, C6 (Creating), there are 2 items questions 39 and 40. These questions ask students to demonstrate creative attitudes and propose solutions in response to Indonesia's cultural and ethnic diversity. These items encourage students not only to understand and apply values but also to develop innovative and solution-oriented perspectives. Overall, the construction of the test items fulfills the principle of cognitive variation, effectively assessing multiple cognitive dimensions aligned with the learning objectives of Pancasila Education. With a balanced distribution across remembering, understanding, applying, analyzing, evaluating, and creating, this assessment is expected to comprehensively explore students' potential and foster the development of critical and creative thinking skills in accordance with curriculum demands..

3.2 Quantitative Analysis

Item Difficulty Level

Item difficulty refers to the extent to which a test item is considered hard or easy. According to Suprananto (2012), item difficulty is defined as the probability of a test item being answered correctly at a certain ability level, typically expressed in the form of an index. Sudjana (2009) explains that the difficulty index ranges from 0.00 to 1.00. This index indicates the difficulty level of a test item: an index of 0.00–0.30 indicates a difficult item, 0.31–0.70 indicates a moderate item, and 0.71–1.00 indicates an easy item. Based on the analysis of the difficulty level of 40 Pancasila Education test items administered in elementary schools in Bengkulu City, the results show that 2 items (5%) fall into the difficult category (items number 24 and 26); 13 items (32.5%) are categorized as moderate (items number 3, 4, 6, 9, 10, 14, 17, 20, 22, 23, 25, 33, and 37); and 25 items (62.5%) are classified as easy (items number 1, 2, 5, 7, 8, 11, 12, 13, 15, 16, 18, 19, 21, 27, 28, 29, 30, 31, 32, 34, 35, 36, 38, 39, and 40).

These results indicate that the majority of the test items tend to be easy, while only a small proportion fall within the high-difficulty level. This finding is crucial for further evaluation to ensure that the composition of test items better reflects a balanced range of difficulty levels, thereby providing a more accurate measure of students' abilities. The percentages are illustrated in the following figure.

Figure 1. Item Difficulty Level Chart

The calculation results based on the data in Figure 1 show a ratio of easy : moderate : difficult items as 25 : 13 : 2. This ratio indicates that the multiple-choice questions have an imbalanced difficulty level distribution, with a dominance of easy questions. Such dominance may reduce the instrument's effectiveness in accurately differentiating students' levels of understanding. Ideally, a test set should have a balanced proportion of difficult, moderate, and easy questions. According to Sudjana (2005), an appropriate distribution might follow a ratio such as 3:4:3 or 2:5:3 for the categories of difficult:moderate:easy items. However, based on the analysis of 40 multiple-choice items from the Pancasila Education subject administered in elementary schools in Bengkulu City, it was found that 25 items (62.5%) were easy, 13 items (32.5%) were moderate, and only 2 items (5%) were difficult, resulting in a ratio of 25:13:2.

This composition shows that the item difficulty level is not yet proportional, as it is heavily dominated by easy questions. Therefore, revisions and adjustments in the construction of test items are necessary to meet recommended standards. A more balanced distribution will help ensure that the assessment instrument can more effectively measure students' abilities comprehensively.

Discriminating Power

According to Daryanto (2007), the discriminating power of a test item refers to its ability to differentiate between high-achieving and low-achieving students, thereby assisting teachers in evaluating student learning more effectively. Arikunto (2003) further explains that the discrimination index of an item indicates how well the item distinguishes between students with high and low levels of ability.

A discrimination index ranging from 0.00 to 0.20 indicates poor discrimination, making the item less effective and in need of revision. A value of 0.21 to 0.40 is considered moderate, still usable though not optimal. Items with a value between 0.41 and 0.70 are regarded as having good discrimination and are suitable for evaluation. Values from 0.71 to 1.00 reflect very good discrimination, meaning the item is highly effective in distinguishing students' ability levels. Conversely, items with negative discrimination indices are considered very poor, as they tend to be answered correctly by lower-ability students more often than by higher-ability ones. Such items should be revised or discarded.

Based on the discrimination index analysis of 40 multiple-choice questions in the Grade VI Pancasila Education assessment, the results show: 2 items (5%) have very poor discrimination (Items 23 and 31), 5 items (12.5%) have poor discrimination (Items 5, 10, 13, 24, and 34), 17 items (42.5%) fall into the moderate category (Items 4, 7, 8, 9, 11, 12, 14, 16, 20, 21, 25, 27, 30, 32, 33, 36, and 37), 16 items (40%) show good discrimination (Items 1, 2, 3, 6, 15, 17, 18, 19, 22, 26, 28, 29, 35, 38, 39, and 40), No items reached the very good category.

These findings indicate that while the majority of items fall within the moderate to good range, several questions exhibit low or even negative discrimination and therefore require revision to better identify differences in students' understanding and performance levels.

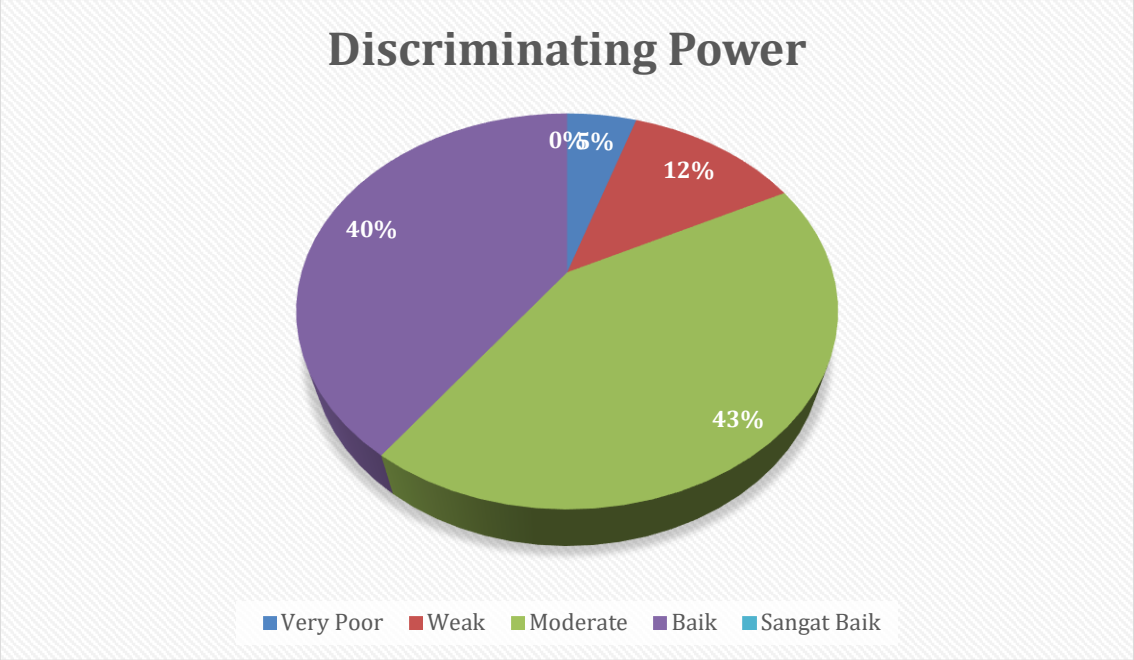

Figure 2. Percentage of Item Discriminating Power

Items categorized as moderate, good, and very good are considered suitable for use in the learning evaluation process because they have adequate ability to distinguish students' levels of ability. Conversely, items that fall into the poor and very poor categories are not recommended for use, as they fail to clearly differentiate between high-ability and low-ability students. Poor items are generally either too easy, so that almost all students answer correctly, or too difficult, resulting in almost no correct responses. Meanwhile, items in the very poor category indicate a mismatch in item quality, where students who should be able to answer correctly fail to do so, while less capable students answer correctly. This may be due to guessing luck or dishonest behaviors such as copying answers from more capable peers. Therefore, it is crucial to review and revise items with low discriminating power to ensure that the assessment instrument accurately reflects students' abilities objectively.

A good test item is one that is answered correctly by more high-ability students than low-ability students, thus effectively distinguishing between the two groups. According to Sudaryono (2012), an item has discriminating power if most high-achieving students answer it correctly while low-achieving students do not. Based on the discriminating power analysis of 40 Grade VI Pancasila Education items, the results show that 2 items (5%) have very poor discrimination, 5 items (12.5%) are poor, 17 items (42.5%) are moderate, and 16 items (40%) are good, with no items reaching very good discrimination. Thus, 57.5% of the items (moderate and good categories) are deemed suitable for evaluation, whereas 17.5% (poor and very poor categories) are considered unsuitable. These findings indicate that the discriminating power of the items is fairly good overall but still requires improvement in some items to optimize the evaluation function.

Distractor Effectiveness

According to Zuldafrial (2012), distractor effectiveness analysis is used to determine whether the distractors function well or not. (Daryanto, 2007:192) states that a distractor is considered effective if it has a strong attraction for test-takers who do not understand the concept or have not mastered the material.

Zuldafrial (2012) further explains that a distractor functions well if it is selected by at least 5% of test participants, meaning that the distractor is capable of attracting students' attention as an answer choice. However, based on the analysis of distractor effectiveness for Grade VI Pancasila Education test items in Elementary Schools in Bengkulu City, it was found that 29 items (72.5%) had ineffective distractors, while only 11 items (27.5%) had distractors that functioned well and effectively. This indicates that the majority of distractors in these items were less able to divert students' choices optimally. This data is illustrated in Figure 3.

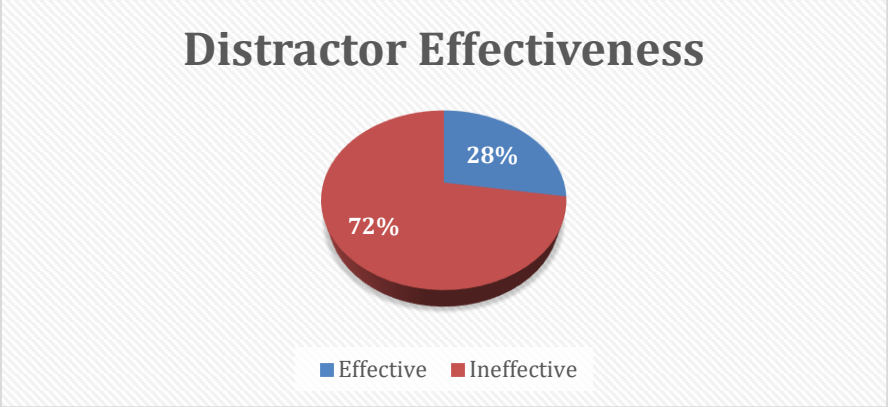

Figure 3. Percentage of Distractor Effectiveness

The questions with ineffective distractors were found in the following items and options: item number 5 (choices C, D), 7 (B, C), 8 (B), 9 (C), 10 (C), 11 (C), 12 (A, C), 14 (C), 15 (A, C, D), 16 (B), 17 (C), 18 (B, C, D), 19 (B, C), 20 (C), 21 (A, B, C), 22 (D), 28 (A, C), 29 (A, D), 30 (A, C), 31 (C), 32 (D), 33 (A, B, D), 34 (D), 35 (C), 36 (A, B, D), 37 (D), 38 (C), 39 (D), and 40 (C). These distractors were deemed ineffective because more high-ability students answered correctly rather than choosing the distractors, indicating that these options failed to attract students to select the wrong answer. Therefore, revisions are needed to improve these distractors so they function properly and enhance the quality of the test as an evaluation tool.

According to Astiti (2017), factors affecting distractor effectiveness include overly easy questions, clues in the question stem leading to the correct answer, and students who have already mastered the material. Multiple-choice questions lacking uniformity in answer options tend to have ineffective distractors. Additionally, when the question stem provides hints, test-takers may guess without careful thinking. Daryanto (2007) adds that effective distractors are acceptable and usable, while ineffective ones should be revised or replaced. Based on the analysis of distractors in the Grade 6 Civic Education test in Bengkulu Elementary Schools, most distractors (72.5%) were ineffective, with only 27.5% functioning well. Therefore, ineffective distractors need to be improved to enhance the overall quality of the test and better divert students' choices.

Validity

Validity is the degree to which a measurement instrument accurately measures what it is intended to measure according to the desired objectives. Based on the validity analysis correlating each item with the total score of the 40 questions, it was found that 21 items (52.5%) were categorized as valid, namely items number 1, 2, 3, 6, 7, 8, 15, 16, 17, 18, 19, 22, 25, 26, 28, 29, 33, 35, 38, 39, and 40. Meanwhile, 19 items (47.5%) were declared invalid, including items number 4, 5, 9, 10, 11, 12, 13, 14, 20, 21, 23, 24, 27, 30, 31, 32, 36, and 37. This indicates that nearly half of the questions used have not met the validity criteria and require revision to improve the quality of the test. The results of the item validity analysis are presented in Figure 4.

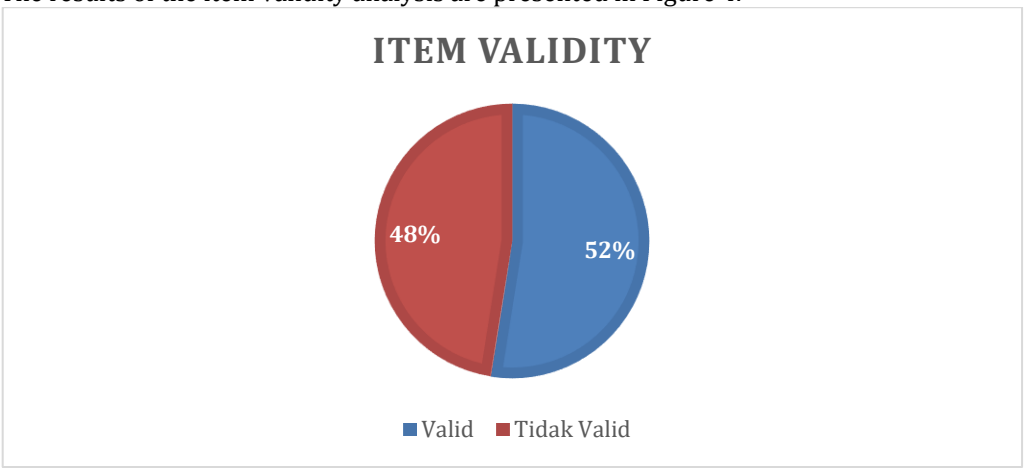

Figure 4. Item Validity

According to Sudaryono (2012), a test or measurement instrument is considered valid if it can perform its function well, meaning it produces measurement results that align with the intended purpose of the measurement. In other words, the measurement results must accurately reflect the actual facts or conditions of the object being measured. A valid item is one that measures exactly what it is supposed to measure. Therefore, the validity of a test must always be linked to its intended use. In the context of achievement tests, the test results should be interpreted so they can be used to evaluate the achievement based on predetermined goals.

Sukardi (2010: 38) states that the validity of an instrument is influenced by several factors, including internal test factors, external test factors, and factors originating from the students themselves. Internal factors include unclear question formulation, which reduces validity, and question difficulty levels that do not correspond to the material taught to students. External factors involve conditions such as insufficient time to complete the test, causing students to answer hurriedly, and cheating during the test, which makes it difficult to distinguish between students who have truly learned and those who have not. Meanwhile, student factors may include anxiety experienced before taking the test due to the testing situation. Questions deemed valid are suitable to be included and used in future evaluations, while invalid questions should be revised or removed.

Reliability

Reliability indicates the consistency or stability of a measurement instrument in assessing the object being measured, so that when the instrument is used repeatedly under the same conditions, the results remain relatively consistent. According to Suprananto (2012: 82), reliability refers to the consistency of test scores from one measurement to another. Reliability values range from 0 to 1, with values closer to 1 indicating higher levels of consistency and accuracy (Suprananto, 2012). The results of the reliability test are presented in Table 3.

Table 3. Reliability Test Results

Reliability Statistics	
Cronbach's Alpha	N of Items
.785	40

Based on the reliability test results of the Semester II Pancasila Education test for sixth-grade students, a Cronbach's Alpha value of 0.785 was obtained from 40 items. This value falls into the high reliability category, indicating that the test has good stability and accuracy in measuring student competencies. Therefore, this test can be considered reliable and suitable to be used as an evaluation instrument in Pancasila Education learning.

Student Understanding Levels

The research results show that students' understanding in answering evaluation questions can be mapped based on Bloom's Taxonomy cognitive domains, which consist of C1 (knowledge), C2 (comprehension), C3 (application), C4 (analysis), C5 (synthesis), and C6 (evaluation). From the total analyzed questions, the largest proportions are in the C2 domain (30%) and C1 domain (22.5%), which fall under the category of Lower Order Thinking Skills (LOTS). Meanwhile, the C3 domain (25%) is categorized as Middle Order Thinking Skills (MOTS), and the remaining C4 (10%), C5 (10%), and C6 (5%) domains fall under Higher Order Thinking Skills (HOTS). Overall, the distribution of questions shows a dominance of LOTS at 52.5%, MOTS at 25%, and HOTS only at 22.5%.

These findings reflect that most of the given questions still focus on students' basic abilities to recall and understand information. However, in the context of 21st-century learning, students are required to think critically, analyze information, and make decisions based on deep reasoning. The small proportion of HOTS questions (C4–C6) indicates the need for improvement in designing questions that challenge students to think more complexly, especially at the elementary school level, which should already encourage higher-order thinking skills. Therefore, these results provide an important basis for teachers and question developers to better balance the proportion of questions according to cognitive levels. Increasing questions that assess analysis, synthesis, and evaluation domains is necessary so that assessments not only measure memorization but also critical thinking

and problem-solving abilities. This will make the evaluation process more representative of the competencies aimed for in learning.

4. Conclusion

Based on the research results, the analyzed Pendidikan Pancasila questions cover various cognitive levels from basic (C1) to higher levels (C6), with a fairly diverse proportion that supports the development of students' critical thinking skills. However, quantitatively, most questions have an easy difficulty level at 62.5%, followed by moderate at 32.5%, and difficult at 5%, which is not yet optimal in measuring the full range of student abilities. The item discrimination index mostly falls into the moderate to good categories, totaling 82.5%, although there are still some questions with low and very poor discrimination that need revision. The effectiveness of distractors remains low, with 72.5% of distractors being ineffective and requiring improvement, while nearly half of the questions (47.5%) do not meet the validity criteria. The reliability of the test bank is high, with a Cronbach's alpha of 0.785, indicating that this instrument can be trusted for learning evaluation. Additionally, the dominance of questions in the LOTS domain at 52.5% shows that the assessment mainly focuses on basic knowledge and comprehension, while questions in the HOTS domain (C4–C6) account for only 22.5%, highlighting the need for increased emphasis on training students in critical, analytical, and creative thinking in line with 21st-century curriculum demands. Revisions and adjustments to the questions are necessary to balance the difficulty levels according to ideal standards for difficult, moderate, and easy items. Questions that are too easy or too difficult should be improved to optimally measure student competencies. Furthermore, questions with low discrimination and ineffective distractors need to be revised or replaced, as distractors must be designed to attract students who have not mastered the material, thereby enhancing the quality of the questions as a more accurate and meaningful evaluation tool.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Adawiah, R., & Maulana, M. F. (2024). Analyzing assessment instruments in Pancasila and citizenship education subjects at Banjarmasin state junior high schools. *Journal of Teaching and Learning Environments*, *1*(1), 18–25.

Anderson, L. W. (2023). *Civic education, citizenship, and democracy*. Education Policy Analysis Archives, 31.

Arikunto, S. (2003). *Prosedur penelitian suatu praktek*. Jakarta: Bina Aksara, 3.

Arikunto, S. (2021). *Dasar-dasar evaluasi pendidikan edisi 3*. Bumi Aksara.

Astiti, K. A. (2017). *Evaluasi pembelajaran*. Penerbit Andi.

Crisp, V., Johnson, M., & Constantinou, F. (2019). A question of quality: Conceptualisations of quality in the context of educational test questions. *Research in Education*, *105*(1), 18–41.

Crittenden, J., & Levine, P. (2018). *Civic Education*. The Stanford encyclopedia of philosophy.

Damayanti, A. M., SH, M. P., Daryono, M. P., & Rayanto, Y. H. (2023). *Evaluasi pembelajaran*. Basya Media Utama.

Daryanto. (2007). *Evaluasi Pendidikan*. Rineka Cipta.

Farida, A. M., & Musyarofah, A. (2021). Validitas dan Reliabilitas dalam Analisis Butir Soal. *Al-Mu'arrib: Jurnal Pendidikan Bahasa Arab, 1*.

Greany, T. (2024). Moral Purpose in Performative Times: Do School leaders' Values Matter? *British Journal of Educational Studies*, *72*(5), 587–606.

Haduong, P., Jeffries, J., Pao, A., Webb, W., Allen, D., & Kidd, D. (2024). Who am I and what do I care about? Supporting civic identity development in civic education. *Education, Citizenship and Social Justice*, *19*(2), 185–201.

Harahap, A. (2024). *Evaluasi Pembelajaran Berbasis Hots Dalam Kurikulum Merdeka*. Penerbit Adab.

Hasibuan, N. H., Safitri, S., & Ariska, I. (2024). Teknik Pengolahan Skor Hasil Evaluasi. *MUDABBIR Journal Research and Education Studies*, *4*(2), 460–475.

Kusumawati, I., Fadillah, N., & Paryanto, P. (2025). Analisis Butir Soal Penilaian Akhir Semester Mata Pelajaran PPKn Menggunakan Program Iteman 3.0 Kelas XI MAN 2 Sleman. *CIVICUS: Pendidikan-Penelitian-Pengabdian Pendidikan Pancasila Dan Kewarganegaraan*, *13*(1), 126–134.

Kwok, Y., & Kwan, C. (2025). Understanding the Factors Affecting Students' Ability to Solve Math Word Problems. *Anatolian Journal of Education, 10*(1), 37–44.
Magdalena, I. (2020). *Evaluasi pembelajaran SD: teori dan praktik*. CV Jejak.
Marsudi, K. E. R., & Sunarso, S. (2019). Contents analysis of the pancasila education and citizenship students' book for high school curriculum 2013. *KnE Social Sciences*, 447–459.
Maryani, I., Prasetyo, Z. K., Wilujeng, I., Purwanti, S., & Fitrianawati, M. (2021). HOTs Multiple Choice and Essay Questions: A Validated Instrument to Measure Higher-Order Thinking Skills of Prospective Teachers. *Journal of Turkish Science Education*, *18*(4), 674-690.
Mbana, M. R. D., Hariadi, F., & Mira, T. D. N. B. (2024). Application of Multimedia Learning for Pancasila and Citizenship Education in SD Inpres Waingapu 3. *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, *3*(3), 864–869.
Prijowuntato, S. W. (2020). *Evaluasi pembelajaran*. Prijowuntato, S. W.
Ramadhan, S., Kusumawati, Y., & Aulia, R. (2024). *Pendidikan dan Pembelajaran Dalam Kurikulum Merdeka di Sekolah Dasar*. Penerbit K-Media.
Ramadhan, W., Malahati, F., Romadhon, K., & Ramadhan, S. (2023). Analisis butir soal tipe multiple choice questions pada penilaian harian sekolah dasar. *Tarbiyah Wa Ta'lim: Jurnal Penelitian Pendidikan Dan Pembelajaran*, *10*(2), 93–105.
Rifana, F., Ramadhan, S., & Putro, K. Z. (2024). Analisis Butir Soal Ulangan Harian Siswa Mata Pelajaran PPKn Kelas IV Menggunakan Rach Model di Madrasah Ibtidaiyah Negeri. *Attadrib: Jurnal Pendidikan Guru Madrasah Ibtidaiyah*, *7*(1), 99–110.
Sudaryono. (2012). *Dasar-Dasar Evaluasi Pembelajaran*. Graha Ilmu.
Sudjana, N. (2005). *Metode Statistika*. Tarsito.
Sudjana, N. (2009). *Penilaian Hasil Proses Belajar Mengajar*. PT. Remaja Rosdakarya.
Sukardi. (2010). *Metodologi Penelitian Pendidikan*. PT Bumi Aksara.
Suprananto. (2012). *Pengukuran dan penilaian pendidikan*. Graha Ilmu.
Susanti, A. (2025). Asesmen Diagnostik Pendidikan Pancasila dalam Pembelajaran Berdiferensiasi pada Siswa Kelas IV SDN 64 Bengkulu Tengah. *Jurnal PGSD: Jurnal Ilmiah Pendidikan Guru Sekolah Dasar*, *18*(1), 63–71.
Susanti, A., Darmansyah, A., Naqsyahbandi, F., & Muktadir, A. (2024). Analyzing student learning style profiles for differentiated learning in merdeka curriculum in elementary schools. *Cendikia: Media Jurnal Ilmiah Pendidikan*, *14*(3), 209–223.
Susanti, A., Darmansyah, A., Tyas, D. N., Hidayat, R., Syahputri, D. O., Wulandari, S., & Rahmasari, A. (2023). The Implementation of Project for Strengthening the Profile of Pancasila Students in the Independent Curriculum for Elementary School Students. *IJECA (International Journal of Education and Curriculum Application)*, *6*(2), 113–122.
Susanti, A., & Muktadir, A. (2025). Pengembangan Media Pembelajaran Gurita Berbasis Karakter untuk Meningkatkan Hasil Belajar Kognitif pada Materi Pengamalan Sila Pancasila Siswa Kelas IV SD. *Jurnal Pembelajaran Dan Pengajaran Pendidikan Dasar*, *8*(1), 28–39.
Sutikno, M. S. (2021). *Strategi pembelajaran*. Penerbit Adab.
Widodo, H. (2021). *Evaluasi pendidikan*. Uad Press.
Zuldafrial. (2012). *Evaluasi Pendidikan & Penelitian Tindakan Kelas*. STAIN Pontianak Press.