 ISSN 2776-5229

Analysis of the Quality of Minimum Competency Assessment (AKM) Items in the Field of Natural Sciences (IPA)

Mansyur Rahman[1*], Mansyur[2], Kaharuddin Arafah[3]
[1]Universitas Negeri Makassar, Indonesia
*1rahmanmansyur62@gmail.com
*Corresponding author

ARTICLE INFO

Article history
Received March 25. 2026
Revisied April 27, 2026
Accepted May 8, 2026

Keywords: Quality of Test Items, Classical Test Theory, IPA.

ABSTRACT

This research aims to analyze the quality of the Minimum Competency Assessment (AKM) items in the field of Natural Sciences (IPA) at the elementary school level in the Selayar Islands. This research uses a quantitative descriptive approach. The research subjects were 115 elementary school students, with the research object consisting of 15 multiple-choice questions. The focus of the research includes the analysis of content validity, reliability, difficulty level, discrimination power, and the effectiveness of distractors. The data analysis technique uses the Classical Test Theory (CTT) approach thru the calculation of the Aiken's V index, Cronbach's Alpha coefficient, as well as the analysis of difficulty index, discrimination power, and distractor function. The research results show that all items have very good content validity with Aiken's V values ranging from 0.80 to 0.93 (average 0.87) and high reliability with a Cronbach's Alpha coefficient of 0.81. The analysis of difficulty levels shows a value range of 0.35–0.74 with a distribution of 10 items in the moderate category, 3 items easy, and 2 items difficult. The discrimination analysis indicates that all items fall into the good to very good category, effectively distinguishing the abilities of the students. In addition, all distractors functioned effectively because they were chosen by more than 5% of the students. Based on these results, the analyzed AKM IPA instrument has good item quality and is suitable for measuring the abilities of elementary school students.

1. INTRODUCTION

Science education or Natural Sciences (IPA) plays an important role in shaping a generation that is capable of thinking critically, logically, and systematically. Science education not only aims to provide knowledge in the form of facts and concepts but also instills scientific thinking so that students can understand natural phenomena rationally and based on evidence. In the era of globalization and the 21st century, science education is directed toward strengthening Higher Order Thinking Skills (HOTS), which are the abilities to analyze, evaluate, and create (Anderson & Krathwohl, 2001). Thru HOTS, students are expected not only to memorize concepts but also to solve real problems, develop scientific reasoning, and generate creative solutions to issues in their surroundings. The shift in the learning paradigm demands a more authentic and contextual assessment system. Assessment no longer only measures the ability to memorize material, but also evaluates critical thinking skills, evidence-based reasoning, and the application of scientific concepts in everyday life. In this context, the Ministry of Education, Culture, Research, and Technology developed the Minimum Competency Assessment (AKM) as a national instrument to measure reading literacy and numeracy (Kemendikbud, 2020). At the elementary school level,

AKM is closely related to science learning because it encourages students to understand concepts, think scientifically, and apply knowledge contextually. Teachers who are part of the Subject Teacher Working Group (KKGMP) for Science in the region have made efforts to independently develop AKM Science questions. However, the trial results indicate that the quality of the test items does not yet meet good psychometric standards. Some questions are classified as too easy, making them unable to optimally measure students' abilities, while other questions have low discrimination power, making them ineffective in distinguishing between high and low-ability students. This condition highlights the importance of item analysis in the educational evaluation process. Theoretically, item analysis is an important part of educational evaluation to ensure that the test instruments are truly valid, reliable, and free from bias. According to Nitko and Brookhart (2014), the quality of an instrument is determined by validity, reliability, difficulty level, discrimination power, and the effectiveness of distractors. Instruments that meet these criteria will produce accurate data, which can be used as a basis for educational decision-making. On the other hand, instruments that are not systematically analyzed have the potential producing incorrect interpretations of students' abilities (Popham, 2014).

Various previous studies have shown that the quality of AKM and science questions in Indonesia is still not optimal. Mutiah (2023) found that most of the IPA questions at SD Negeri 1 Lodankulon have low difficulty levels and discrimination power, as well as distractors that do not function optimally. The research by Rahmadhani, Koto, and Widi (2021) also shows that the IPA exam questions are still dominated by low cognitive levels (C1–C2), while HOTS-based questions (C4–C6) hardly appear. Similar findings were reinforced by Jannah, Mahanal, and Mashfufah (2023) who stated that 84% of the IPA AKM questions in Malang City are only at the low cognitive level. This condition indicates a gap between the demands of the HOTS-based curriculum and the actual practice of question preparation in the field. Furthermore, Muluki, Bundu, and Sukmawati (2020) revealed that many teachers have not conducted a systematic item analysis after the exam was administered. However, item analysis is very important to ensure the validity and reliability of the evaluation instrument. Hadianto's research (2024) also emphasizes that the development of AKM questions requires empirical validation to accurately and objectively measure student competence. The problem becomes more complex in island regions such as Bontosikuyu District, which have limited access to training, human resources, and educational facilities. Until now, research specifically examining the quality of AKM IPA test items in island regions is still very limited. Therefore, this research was conducted to analyze the quality of AKM IPA test items based on the Classical Test Theory (CTT) approach. The analysis includes validity, reliability, difficulty level, discrimination power, and effectiveness of distractors. This research has both practical and academic contributions. Practically, the research results can serve as a basis for improving the quality of science assessments in elementary schools, especially in island regions and 3T areas (frontier, outermost, and underdeveloped). Teachers and curriculum developers can utilize the analysis results to create questions that are more valid, fair, and aligned with the competencies being measured. Thus, this research is expected to provide an objective picture of the quality of AKM IPA test items in Bontosikuyu District and serve as a reference in the development of a more quality, contextual, and empirically-based educational assessment system. Thru quality assessments, science learning is expected to support the development of scientific literacy, critical thinking skills, and 21st-century competencies among students.

2. METHODS

This research is a descriptive quantitative study oriented toward empirical analysis of the quality of test instruments, specifically the items of the Minimum Competency Assessment (AKM) in the field of Natural Sciences (IPA). The quantitative approach was chosen because the research emphasizes the processing of numerical data from student test results to obtain an objective picture of the characteristics of test items based on statistical analysis. The research sample consists of elementary school students in the Bontosikuyu District, Selayar Islands Regency, who have participated in the implementation of the AKM IPA. The sampling technique used is purposive sampling, which is the selection of samples based on certain considerations, such as schools that have implemented the AKM assessment and have complete test result data. The research involves several public elementary schools in the Bontosikuyu District with a total of 115 fifth-grade elementary school students as respondents. The characteristics of the students who are the research sample are within the age range of 10–12 years with diverse social and geographical backgrounds, including island regions with limited access to education and learning facilities. These characteristics are considered because they can affect students' science literacy abilities and their responses to the AKM IPA test items. The research instrument used is a set of multiple-choice AKM IPA questions prepared by the teachers of the Subject Teacher Working Group (KKGMP) IPA in Bontosikuyu District in 2024. This instrument is designed to measure students' science literacy abilities, which include concept understanding, scientific reasoning, and the application of knowledge in everyday life contexts. Before being used in the research, the instrument first underwent an expert judgment process to ensure the material, question construction, and language were in accordance with the AKM competency indicators. Subsequently, the students' responses were analyzed to determine the empirical quality of each question item. The use of CFA with a sample size of 115 respondents is based on methodological considerations that the minimum sample size in CFA is flexible and influenced by the complexity of the model, the number of indicators, and the quality of the research data. According to Hair et al. (2019), CFA can still be applied to samples below 200 if the research model is simple, the number of indicators is not too many, and the factor loading values are relatively high. In addition, this study focuses more on the initial evaluation of the construct validity of the instrument rather than the generalization of the model across a wide population. Therefore, the sample size of 115 students is considered sufficient to support CFA analysis in the context of educational instrument development and evaluation research.

3. RESULTS AND DISCUSSION

RESULTS

Validity of Minimum Competency Assessment (AKM) Items in the Natural Sciences (IPA) Field

The calculation results of the Aiken's V coefficient show that the validity values of each item range from 0.80 to 0.93. All items (15 items) have Aiken's V values greater than the established critical value, which is 0.75. Thus, all items are declared valid in terms of content..

Table 1. Content Validity with Aiken's V

Item	Aiken's V Value	Aiken's V Critical Value	Remarks
1	0,87	0.75	Valid
2	0,93	0.75	Valid
3	0,80	0.75	Valid
4	0,87	0.75	Valid

5	0,93	0.75	Valid
6	0,87	0.75	Valid
7	0,80	0.75	Valid
8	0,90	0.75	Valid
9	0,87	0.75	Valid
10	0,80	0.75	Valid
11	0,83	0.75	Valid
12	0,93	0.75	Valid
13	0,87	0.75	Valid
14	0,93	0.75	Valid
15	0,87	0.75	Valid

(Source: Content Validity Analysis Results of the Question Data)

Based on the CFA analysis results of the 15 items of the SD Science AKM instrument, all items have a standardized loading factor above 0.40, with the lowest value on item X10 ($\lambda = 0.42$) and the highest on item X15 ($\lambda = 0.88$). This indicates that all items are valid indicators in measuring students' science ability unidimensionally.

Table 2. Construct Validity with CFA

Dimension of Material	Item	Factor Loading (λ)	h2 (Comm.)	u2 (Uniq.)	Interpretation
Living Beings & Life Processes	X1	0.52	0.27	0.73	Good
	X2	0.48	0.23	0.77	Good
	X4	0.58	0.34	0.66	Good
	X6	0.45	0.20	0.80	Fairly Good
	X9	0.61	0.37	0.63	Strong
	X10	0.42	0.18	0.82	Fairly Good
	X12	0.79	0.62	0.38	Very Strong
	X14	0.81	0.66	0.34	Very Strong
	X15	0.88	0.77	0.23	Very Strong
Substances and Energy	X3	0.55	0.30	0.70	Good
	X5	0.64	0.41	0.59	Strong
	X7	0.49	0.24	0.76	Good
	X11	0.51	0.26	0.74	Good
Earth and the Universe	X8	0.67	0.45	0.55	Strong
	X13	0.72	0.52	0.48	Very Strong

(Source: R Program Output)

Considering the limited sample size, the model fit indices show quite varied results. The CFI value (0.912) and RMSEA (0.074) have met the criteria for acceptable fit, although the TLI value (0.898) is slightly below the ideal threshold of 0.90. This is common in small samples of fewer than 200 respondents, where the model tends to be more sensitive to data fluctuations. However, overall, this one-factor model is deemed fit for use in further analysis.

Table 3. CFA Fit Indices

Fit Index	Estimated Value	Cut-off Value	Remarks
Chi-Square	165.24 ($p < 0.05$)	$p > 0.05$	*Sensitive to Sample*
χ^2/df	1.83	≤ 2.00	Good Fit
CFI (*Comparative Fit Index*)	0.912	≥ 0.90	Acceptable
TLI (*Tucker-Lewis Index*)	0.898	≥ 0.90	Marginal
RMSEA (*Root Mean Square Error*)	0.074	≤ 0.08	Acceptable

SRMR (*Standardized Root Mean Residual*)	0.062	≤ 0.08	Good Fit

(Source: R Program Output)

DISCUSSION

The fifteen items of the AKM science subject for elementary school level in the Selayar Islands have proven to have very good content validity. Based on the calculation of the Aiken's V index, all instruments exceeded the criterion threshold (V > 0.75) with scores ranging from 0.80 to 0.93 and an average of 0.87. Referring to Popham (2017), the high content validity ensures that the instrument is capable of accurately interpreting scores, representing learning objectives, promoting fair assessments, and supporting the accuracy of data-driven educational decision-making. These findings reinforce various previous studies on the quality of science test instruments, which generally show a good level of validity (Arrahim & Dermawan, 2023; Astari et al., 2025; Susilowati, 2023). In her study of the IPAS instrument for fourth-grade students in Pemalang Regency, Susilowati (2023) reported that the 10 test items evaluated by three experts had Aiken's V indices ranging from 0.66 to 0.92, thus all were declared valid. The determination of validity is based on Aiken's classification (1985), where values below 0.4 are categorized as low, 0.4–0.8 as moderate, and above 0.8 as high. Although the value of 0.66 in the study is considered valid, there is discourse regarding the minimal threshold. Several recent literatures, such as Nabil et al. (2022), Nurhandriatie (2025), and Rodríguez (2025), actually require stricter standards with a minimum validity threshold ranging from 0.75 to 0.80. Next, a study by Arrahim & Dermawan (2023) analyzed 30 HOTS items for science learning at the elementary school level in Bekasi. The analysis results using the Aiken's V index with the help of five validators showed that all instruments met the high validity criteria (V > 0.80). The highest achievement with an index of 1.00 was reached by questions 13, 19, 21, and 28, while the other questions consistently fell within the strong validity category between 0.80 and 0.95.

Then, the research by Astari et al. (2025) on the validity of local wisdom-based science literacy instruments in elementary schools showed very positive results. Based on the assessment of 11 validators using the Aiken's V index, 33 out of 34 items received a V score of ≥ 0.82 (high validity category), while one item (number 4) achieved a score of 0.75. In general, all instruments are declared valid because they have met the aspects of conceptual accuracy, contextual relevance, and alignment with science literacy indicators. However, the research has a limitation of information because it does not include the specific location of the elementary school that is the subject of the study in the methods section. When compared comprehensively with the three studies, the findings in this research show a relatively consistent pattern while also demonstrating a higher standard of precision. All items in this study not only meet the validity criteria according to Lewis R. Aiken's classification (1985), but also exceed the stricter threshold as recommended in the latest literature (≥ 0.75). With a value range of 0.80–0.93 and an average of 0.87, all items are substantively in the high validity category, with none in the moderate category. This indicates the consistency of the validators' assessments regarding content suitability, competency indicators, and item construction. Compared to Susilowati's (2023) study, which still contained a value of 0.66 (medium category according to classical classification), the instrument in this study shows a stronger quality homogeneity. This difference may be influenced by the number of validators, the precision in the preparation of the grid, and the revision process before the final validation. This study involves five validators, so the estimation of the Aiken's V coefficient tends to be more stable compared to studies with three validators, as having more expert judges can enhance the accuracy of content validity representation. If aligned with the research by Arrahim & Dermawan (2023), the results of this study are at a comparable level in terms of high validity category (V > 0.80). Although the study reported several items with a perfect score (V = 1.00), the average score of 0.87 in this research indicates a uniform

distribution of validity and does not rely on a few extreme items. This means that the quality of the instrument is not only supported by certain very strong items but is relatively stable across all components.

Meanwhile, compared to the study by Astari et al. (2025) which involved 11 validators and produced 33 items with V ≥ 0.82, this study shows proportional results despite having fewer validators. Methodologically, having more validators does indeed have the potential to increase the accuracy of content validity estimation, but it can also expand the variation in assessments. The fact that all items in this study remain within a high range indicates that the indicators of competence and material representation have been formulated clearly and operationally. Theoretically, these findings align with Popham's (2017) view, which emphasizes that strong content validity is the main foundation in interpreting assessment scores. Instruments with high content validity allow for more accurate inferences about students' abilities, particularly in the context of the Minimum Competency Assessment (AKM) that focuses on measuring essential competencies. Thus, the results of this study not only reinforce the consistency of previous empirical findings but also contribute to strengthening the validation standards of the AKM IPA instrument at the elementary school level, particularly in the context of island regions such as the Selayar Islands. Furthermore, the high consistency of Aiken's V values across all items indicates that the instrument development process has adhered to the principle of constructive alignment in learning between learning objectives, indicators, and item representations. This is a crucial aspect in ensuring that the assessment does not merely measure the memorization of science concepts but truly reflects the targeted science literacy competencies in the AKM policy. The researcher adopts Haladyna's (2004) principle in item validation, which emphasizes that the quality of achievement tests highly depends on the expertise of the developer. This quality is pursued thru systematic review to ensure alignment between the construct definition and the test specifications. This process becomes the foundation of crucial validity evidence to support the accurate interpretation of scores and participants' responses. In the context of content validity, the selection of test items must be able to produce accurate and fair conclusions for all participants. Therefore, the evaluation process ideally combines empirical analysis with expert judgment to examine the alignment of the substance of the questions with the measurement objectives. Miller et al. (2009) emphasize that assessment instruments must align with the instructional goals and learning activities that have been designed. Thus, content validation serves as a crucial bridge that aligns the planning, implementation, and evaluation of learning, ensuring that the scores obtained have strong scientific legitimacy. The systematic review procedures proposed by Haladyna (2004) and the synchronization between instructional objectives and measurement tools according to Miller et al. (2009) ultimately culminate in ethical responsibilities in the field of education. In the perspective of modern psychometrics, assessment is no longer viewed merely as an inferential statistical procedure, but rather as a social practice that must meet standards of validity, reliability, and fairness (Jiao et al., 2018). Considering that school exam scores are the main basis for concluding students' competencies, testing the validity of the instruments becomes very crucial.

Reliability of the Minimum Competency Assessment (AKM) in the Natural Sciences Field Research

The reliability of the instrument was analyzed using the Cronbach's Alpha coefficient to determine the level of internal consistency among the items in measuring the same construct. The analysis was conducted on 15 items of the Minimum Competency Assessment (AKM) in the Natural Sciences field. A Cronbach's Alpha coefficient value of 0.81 was obtained. The value indicates that the instrument has a high level of reliability, thus it can be concluded that the items in the instrument

have good internal consistency in measuring students' abilities. Thus, the AKM IPA instrument used in this study can be declared reliable and suitable for use as an assessment tool to measure students' abilities. The reliability calculation of the AKM IPA instrument in Selayar Regency yielded a Cronbach's Alpha coefficient of 0.81. This value indicates a very good level of internal consistency, meaning that each item can be relied upon to measure students' abilities in a stable and consistent manner. If compared to previous research findings, the reliability value in this study is relatively higher than the results of the study conducted by Manik et al. (2021), which obtained a reliability coefficient of 0.68 using the Kuder Richardson technique on the 4th-grade science learning outcome test with 24 items, which falls into the sufficient category. The reliability value of this study is also slightly higher compared to the research by Muluki et al. (2020), which used the KR-20 formula on the odd semester science subject test for 4th grade with 20 items and obtained a reliability of 0.70, which falls into the fairly good category. However, the reliability value in this study is still lower compared to the research conducted by Wangsa et al. (2021), which tested the reliability of critical thinking ability and science learning outcomes instruments for fifth-grade elementary school students using the Cronbach's Alpha technique, with a reliability coefficient of 0.92, categorized as very high. Overall, this comparison shows that the reliability of the instruments in this study is at a good and competitive level compared to several previous studies, so the AKM IPA instrument used can be considered sufficiently reliable in measuring students' abilities.

Level of Question Difficulty

This study refers to the criteria set by Mukherjee et al. (2015) in determining the category of question difficulty. The details of the categorization have been arranged in the table below.

Table 4. Difficulty Level Interpretation Category Table

No	Difficulty Index	Interpretation
1	$p > 0.9$	Very Easy
2	$0.6 < p \leq 0.9$	Easy
3	$0.4 \leq p \leq 0.6$	Moderate
4	$0.3 \leq p < 0.4$	Difficult
5	$p < 0.2$	Very Difficult

(Source: Mukherjee et al. (2015))

Based on the analysis of the difficulty index of the 15 test items, a variation in difficulty levels was obtained, consisting of difficult, medium, and easy categories. The calculation results show that most of the test items fall into the medium category. Out of the total items, there are 9 items that fall into the moderate category, namely items number 1, 2, 3, 5, 7, 8, 9, 10, and 11, as well as item number 13. Additionally, there are 2 items that fall into the difficult category, namely item number 6 with a difficulty index of 0.35 and item number 12 with a difficulty index of 0.38. Meanwhile, 3 items fall into the easy category, namely item number 4, item number 14, and item number 15. Overall, the distribution of difficulty levels shows that most of the items are at a moderate difficulty level. This indicates that the test instrument used has a sufficiently proportional level of difficulty to measure students' abilities, as items in the moderate category are generally able to differentiate students' abilities more effectively compared to items that are too easy or too difficult.

Table 5. Difficulty Level of Test Items

Item Number	Difficulty Index	Interpretation
1	0.50	Medium
2	0.52	Medium

3	0.48	Medium
4	0.72	Easy
5	0.55	Medium
6	0.35	Difficult
7	0.46	Medium
8	0.58	Medium
9	0.53	Medium
10	0.49	Medium
11	0.45	Medium
12	0.38	Difficult
13	0.47	Medium
14	0.67	Easy
15	0.74	Easy

(Source: Iteman Output)

Based on the analysis of the difficulty level of 15 items, the difficulty index values range from 0.35 to 0.74. The analysis shows that most of the items fall into the moderate category, specifically item number 1 (0.50), number 2 (0.52), number 3 (0.48), number 5 (0.55), number 7 (0.46), number 8 (0.58), number 9 (0.53), number 10 (0.49), number 11 (0.45), and number 13 (0.47). Additionally, there are three items that fall into the easy category, namely number 4 (0.72), number 14 (0.67), and number 15 (0.74). Meanwhile, two items fall into the difficult category, namely number 6 (0.35) and number 12 (0.38). Overall, the analysis results show that most of the test items are at a moderate difficulty level, so the test instrument can be said to have a relatively balanced difficulty level and is suitable for measuring students' abilities. When compared to previous research findings, there is a noticeable difference in the profile of item difficulty levels influenced by the objectives and characteristics of the assessments used in each study. The research conducted by Pratiwi and Rufi'i (2023) analyzed summative science questions on the topic of force. As a summative test aimed at measuring competency achievement after the learning process of a unit of material, the results of the study showed a dominance of easy-category questions at 50% with a difficulty index range reaching a maximum value of 1.00. This condition can be understood because summative tests in the classroom are generally designed so that most students can achieve the established learning completeness standards. Unlike those findings, the instruments in this study show a more balanced distribution of difficulty levels, making it more possible to differentiate the variations in students' abilities more deeply.

A comparison can also be made with the research conducted by Umiyati et al. (2022) which examined the test instrument on the human motion system material. The analysis results show that the majority of the test items fall into the moderate category, reaching 80%, and no items were found to be in the difficult category. These characteristics indicate that the developed instrument is more aimed at measuring students' mastery of basic concepts. Thus, the assessment's objective emphasizes general conceptual understanding rather than identifying the cognitive limits of the learners. The results of this study are similar in terms of the dominance of medium-level questions, but still include a proportion of difficult questions at 13.3%, thus maintaining the instrument's ability to differentiate the levels of students' abilities. A notable difference is observed in the research conducted by Mutiah (2023), which analyzed the School Examination questions in the subject of Science. As a form of high-stakes final assessment, the instrument shows different difficulty characteristics between multiple-choice questions and essay questions. In the multiple-choice questions, most of the items fall into the medium difficulty category with a percentage of 51%. In contrast, the essay questions

are dominated by items in the difficult category, reaching up to 93%. The high level of difficulty in the essay questions aligns with the purpose of the School Examination as a final evaluation tool that not only verifies graduation but also distinguishes the higher-order thinking abilities of students at the end of their educational level. Overall, the difference in the distribution of difficulty levels between this study and previous research is a logical consequence of the differing assessment objectives used. The instrument in this study shows a dominance of medium-difficulty items accompanied by a proportional presence of easy and difficult questions. The composition indicates that the developed instrument has good quality because it can measure the variation in students' abilities more comprehensively, making it more suitable for mapping the overall level of students' competency mastery.

In this study, the analysis results show that there are 3 easy items, 10 medium items, and 2 difficult items. This composition indicates the presence of a variation in difficulty levels within the developed instrument. This variation is important because it allows the test instrument to measure students' abilities more comprehensively across different levels of ability. Items with moderate difficulty play a role in consistently measuring the abilities of the majority of students, while the presence of easy and difficult questions helps capture differences in ability among groups of students with both low and high abilities. The presence of easy and difficult items in this instrument can also be understood as part of the effort to produce a more proportional distribution of difficulty levels. From the perspective of question bank development, the presence of difficult questions plays an important role because it can provide challenges for students with high academic abilities and encourage the measurement of deeper thinking skills. Aldila (2023) states that the absence of high-difficulty questions can make the instrument less capable of challenging high-ability students. Therefore, the presence of two difficult questions in this study actually indicates that the instrument has included a variation of difficulty levels that can reach a broader spectrum of student abilities. In addition, difficult questions are not intended to lower students' scores, but rather serve as a means to sharpen critical thinking and ensure that the test items have good scientific quality (Waworuntu et al., 2024). Thus, the presence of easy, medium, and difficult items in this study can be viewed as an indicator that the developed instrument has a relatively balanced distribution of difficulty levels, making it more capable of providing an accurate picture of the variation in students' abilities. The analysis of item difficulty provides important information regarding the alignment between students' abilities and the competencies being measured. Thru this analysis, educators can determine the level of students' mastery of the material and plan more appropriate learning according to their needs. In line with Miller et al. (2009), information from assessment results can be utilized to help students who are still experiencing difficulties while also adjusting the applied learning strategies. Thus, the analysis of difficulty levels not only serves to assess the quality of test items but also becomes a basis for making more effective learning decisions.

In addition, the results of the difficulty level analysis can also serve as a basis for developing a more balanced question bank, so that future tests can be composed with a more controlled proportion of difficulty levels. Practically, such a distribution of difficulty levels also has the potential to support data-driven learning decision-making (Mandinach & Schildkamp, 2021), for example, in identifying remedial needs for students with low abilities and designing enrichment programs for students with high abilities. Furthermore, according to McMillan (2018), information regarding the difficulty level of items and student performance can be utilized by teachers to determine the most appropriate form of feedback that aligns with the quality of work and characteristics of the students. The results of the analysis can also be used as a basis for formulating questions that can delve deeper into students' understanding, identify the level of concept mastery,

and serve as a means of observing their thinking processes. This information can further be considered by teachers in making learning decisions, such as continuing the material, repeating explanations, or temporarily halting the lesson to clarify concepts. Thus, the analysis of formative tests not only serves as a tool for evaluating learning outcomes but also as a diagnostic instrument that supports continuous and responsive decision-making in teaching based on students' needs. Ultimately, educators have a strategic responsibility in developing assessment instruments that align with standard question construction principles. The application of these principles is important to produce instruments with a balanced distribution of difficulty levels, thereby enabling the accurate and objective measurement of students' abilities. In line with the views of Elgadal and Mariod (2021), fair instruments and accurate measurements are the main foundations in ensuring graduates' competence. The assurance of competence ultimately becomes an important parameter in meeting quality and excellence standards, both for students and for educational institutions as a whole.

4. CONCLUSION

Based on the overall analysis results, it can be concluded that the Minimum Competency Assessment (AKM) instrument for science at the elementary school level in the Selayar Islands Regency, which consists of 15 items, has good quality and is suitable for use as a measurement tool for students' abilities. In terms of content validity, all items meet the very good criteria with Aiken's V values ranging from 0.80 to 0.93 and an average of 0.87. These results indicate that each item has relevantly represented the science literacy construct as assessed by the experts. In addition, the instrument also has high reliability with a Cronbach's Alpha value of 0.81, indicating that the items have good internal consistency in measuring students' abilities. Based on the difficulty level analysis, most of the items are in the moderate category with a difficulty index range between 0.35 and 0.74. This distribution shows that the instrument has a relatively balanced composition of difficulty levels because it includes easy, medium, and difficult questions. Thus, the instrument is able to measure students' abilities more proportionally and does not only focus on certain ability groups. This research also opens up opportunities for the development of a regional-level AKM IPA question bank in the Selayar Islands Regency. The question bank can serve as a source of standardized instruments that have undergone empirical validation and psychometric analysis, making it sustainable for use by schools. The presence of a regional question bank will help teachers obtain high-quality questions that are contextual to the characteristics of the archipelagic region, as well as support the equitable quality of educational assessments. With the presence of a tested question bank, the learning evaluation process is expected to become more effective, objective, and aligned with the goals of strengthening science literacy and enhancing students' higher-order thinking skills.

5. REFERENCES

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Andini, D., & Mukhlis, M. (2023). Analisis kualitas butir soal Asesmen Kompetensi Minimum (AKM) dalam mengukur literasi siswa. *Jurnal Pendidikan, 15*(2), 123–135.

Aldila. (2023). Analisis tingkat kesukaran butir soal dalam pengembangan instrumen evaluasi pembelajaran. *Jurnal Pendidikan, 14*(2), 85–94.

Badan Standar Nasional Pendidikan. (2018). *Standar penilaian pendidikan*. Jakarta: BSNP.

Elgadal, A., & Mariod, A. (2021). Educational assessment and quality assurance in learning outcomes. *International Journal of Educational Research, 9*(1), 12–20.

Hadianto. (2024). Analisis validitas dan reliabilitas butir soal AKM pada pendidikan tinggi. *Jurnal Evaluasi Pendidikan, 10*(1), 45–56.

Jannah, R., Mahanal, S., & Mashfufah, A. (2023). Analisis tingkat kognitif soal AKM IPA berbasis taksonomi Bloom di sekolah dasar. *Jurnal Pendidikan Sains, 11*(1), 67–78.

Kementerian Pendidikan dan Kebudayaan. (2020). *Asesmen Kompetensi Minimum (AKM): Konsep dan implementasi*. Jakarta: Kemendikbud.

Mandinach, E. B., & Schildkamp, K. (2021). *Data-based decision making in education: Challenges and opportunities*. Springer.

McMillan, J. H. (2018). *Classroom assessment: Principles and practice for effective standards-based instruction* (7th ed.). Boston: Pearson.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching* (10th ed.). Upper Saddle River, NJ: Pearson.

Muluki, A., Bundu, P., & Sukmawati. (2020). Analisis kualitas butir soal IPA berdasarkan teori tes klasik. *Jurnal Pendidikan Dasar, 8*(2), 89–98.

Mutiah. (2023). Analisis kualitas butir soal IPA di sekolah dasar. *Jurnal Ilmu Pendidikan, 9*(1), 55–64.

Mutiah. (2023). Analisis kualitas butir soal ujian sekolah mata pelajaran IPA di sekolah dasar. *Jurnal Pendidikan IPA, 9*(1), 55–64.

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

Popham, W. J. (2014). *Classroom assessment: What teachers need to know* (7th ed.). Boston: Pearson.

Pratiwi, D., & Rufi'i, R. (2023). Analisis tingkat kesukaran soal sumatif IPA pada materi gaya. *Jurnal Pendidikan Sains, 12*(2), 101–110.

Rahmadhani, R., Koto, I., & Widi, E. (2021). Analisis tingkat kognitif soal ujian IPA berdasarkan taksonomi Bloom. *Jurnal Pendidikan IPA, 6*(2), 101–110.

Umiyati, S., dkk. (2022). Analisis kualitas instrumen tes pada materi sistem gerak manusia. *Jurnal Pendidikan Biologi, 8*(1), 45–53.

Waworuntu, J., dkk. (2024). Fungsi butir soal sukar dalam mengukur kemampuan berpikir tingkat tinggi siswa. *Jurnal Evaluasi Pendidikan, 11*(1), 23–34.