PAEDAGOGIA
Jurnal Penelitian Pendidikan
http://jurnal.uns.ac.id/paedagogia

DOI : 10.20961/Paedagogia.v29i1.115663
Vol. 29, No. 1, 2026, pp. 88-98
p-ISSN 0126-4109
e-ISSN 2549-6670

Developing an Ethnoscience-Based AKM instrument: Evidence of Content Validity and Item Quality

Sukmaningrum Latifah Oktaviani*, Sri Yamtinah, Budi Hastuti
Departement of Chemistry Education, Universitas Sebelas Maret, Surakarta, Indonesia.

Keywords: *AKM, Test, Developing, Item Quality*

Article history
Received: 6 February 2026
Revised: 20 February 2026
Accepted: 27 February 2026
Published: 28 february 2026

*Corresponding Author Email:
sukmalatifah23@student.uns.ac.id

DOI: 10.20961/paedagogia.v29i1.115663

Abstract: This study uses a descriptive quantitative approach that aims to describe and analyze the quality of Minimum Competency Assessment (AKM) questions developed in integrated ethnoscience learning with a focus on scientific literacy and numeracy literacy in the topic of salt hydrolysis. There are various question formats, namely multiple choice, complex multiple choice, matching, short answer (fill in the blank), and true-false. The item quality analysis in this study involved: difficulty level testing, discrimination power testing, and instrument reliability testing. The difficulty level testing was conducted to determine the proportion of students who were able to answer each item correctly. The discrimination power testing aimed to determine the ability of each item to distinguish between high-ability and low-ability students. Meanwhile, the reliability testing was conducted to assess the overall internal consistency of the test instrument in measuring students' scientific literacy and numeracy skills.The variety of question formats is designed to accommodate the characteristics of AKM which emphasizes higher-order thinking skills, conceptual understanding, and the application of knowledge in the context of everyday life based on ethnoscience.The analysis results show that the majority of test items fall into the moderate difficulty category, indicating that they are neither too easy nor too difficult for students. Furthermore, the discriminatory power of the test items is in the good category, meaning each item effectively differentiates students with different ability levels. High instrument reliability indicates that the test instrument has good consistency and can provide stable and reliable measurement results.

How to cite: Nastiti, I. A., Noviabahari, J. L. & Prakosha, D. (2026). Initial Need Analysis of Applied English for Students of English Study Program in the Hospitality Field. *PAEDAGOGIA*, 29(1), 88-98. DOI: 10.20961/paedagogia.v29i1.115663

INTRODUCTION

One important component of learning that requires attention is assessment (Yamtinah et al., 2017). Assessment is a step taken to measure student learning outcomes. Conducting an assessment requires an instrument (tool) so that the results obtained can be objective and provide an accurate measurement of student learning outcomes, whether using tests or non-tests (Asrul et al., 2014). In 2019, the Minister of Education and Culture officially announced that the National Examination (UN) used to assess student graduation would be abolished and replaced with the Minimum Competency Assessment (AKM) (Ulyah et al., 2021). Some of the reasons that led to the replacement of the National Examination with the Minimum Competency Assessment include; the questions in the National Examination mostly measure low-level thinking competencies which are not in accordance with the educational objectives of developing high-level thinking skills and other competencies relevant to the 21st Century, then the National Examination does not provide enough encouragement to teachers to provide effective teaching methods for the development of high-level thinking skills, the competency assessment replacing the National Examination is designed to provide a strong impetus for innovative teaching and is oriented towards the development of reasoning, not just memorization (Ministry of Education and Culture, 2020).

The assessment used to replace the National Examination is the Minimum Competency Assessment (AKM), which assesses the basic competencies required for all students to develop their capacities and participate positively in society (Asrijanty, 2020). This student assessment is not based on the subjects covered in the national exam, but rather incorporates students' literacy and numeracy

competencies (Purwati et al., 2021). The AKM questions can be multiple-choice, complex multiple-choice, matching, short answer, and true-false statements (Asrijanty, 2020). The AKM is used to measure two basic competencies: numeracy and literacy.

PISA results show that Indonesian students' scientific literacy and numeracy achievements are still below the international average. This situation indicates the need for systematic efforts to improve the quality of learning and assessments that can measure and train higher-order thinking skills. Following up on this, the Indonesian government implemented the Minimum Competency Assessment (AKM) as part of the National Assessment, which is oriented towards measuring reading literacy, numeracy literacy, and science literacy in a contextual and applicable manner. Indonesia ranked 64th out of 72 countries with an average score of 403 for scientific literacy, far below the international average in 2015 (OECD, 2015). Then in 2018, Indonesia scored 396, with only 34% of Indonesian students having a minimum level of scientific competence or more (OECD, 2018). In 2022, a score of 383 for scientific literacy means that Indonesian students' abilities are still far below the international PISA average, which is generally around 485 (OECD, 2023).

In its implementation, the AKM instrument used in schools is still dominated by general contexts that are not close to the real experiences of students (Sumarni, et.al, 2017). As a result, students have difficulty in understanding the problems presented, so that the literacy skills measured do not fully reflect their true potential. Therefore, it is necessary to develop an AKM instrument that is not only statistically valid and reliable, but also contextually meaningful. One approach that is considered relevant to address this problem is the integration of ethnoscience in the development of AKM instruments (Dewi et.al., 2021). Ethnoscience is not merely an effort to incorporate local wisdom into science learning, but an approach that recognizes local cultural practices as a legitimate knowledge system for understanding scientific phenomena (Dewi et al., 2019). It positions community traditions, indigenous technologies, and cultural activities as meaningful contexts for constructing scientific and numeracy concepts. In this perspective, knowledge is not viewed as culturally neutral, but as embedded within social and cultural experiences.

In the context of the Minimum Competency Assessment (AKM), previous literacy and numeracy assessment studies have generally employed broad real-life contexts such as environmental issues, economic activities, or daily quantitative problems. Although these contexts are relevant, they tend to be culturally general and do not explicitly integrate specific local knowledge systems. Consequently, the assessment may measure cognitive reasoning skills but does not optimally connect students' cultural experiences with scientific meaning-making. Sudarmin (2014) argues that the integration of ethnoscience in learning can enhance scientific literacy by making concepts more contextual and culturally meaningful. However, limited studies have explicitly integrated ethnoscience into AKM-based assessment instruments, particularly in the domain of science literacy and numeracy, while simultaneously examining the psychometric quality of the developed items.

This study offers several novel contributions. First, it proposes a conceptual framework that systematically integrates ethnoscience principles into the AKM assessment structure, linking local cultural practices with science literacy and numeracy indicators. Second, it develops and validates an AKM-based instrument embedded in specific local cultural contexts rather than general real-life situations. Third, it provides a comprehensive analysis of the instrument quality, including content validity, construct alignment, and item characteristics. These aspects distinguish this research from previous AKM or literacy assessment studies that primarily used culturally contexts.

METHOD

This study used a descriptive quantitative approach to analyze the quality of the Minimum Competency Assessment (AKM) questions for integrated ethnoscience science and numeracy literacy on salt hydrolysis. The research instrument consisted of AKM questions designed to measure scientific literacy (explaining scientific phenomena, interpreting data and scientific evidence) and numeracy literacy (using numbers, calculations, and interpreting data). The questions were developed based on high school chemistry learning outcomes in salt hydrolysis and AKM indicators.

Ethnoscience integration was achieved by linking the concept of salt hydrolysis to the context of

local wisdom and everyday life, such as traditional salt making, the use of MSG in timlo (a traditional Indonesian dish made from rice flour), the use of leavening agents, natural cough remedies, and the use of soap nuts as a soap for washing batik cloth.

The researchers developed a total of 25 test items designed to measure students' scientific literacy and numeracy skills within the context of ethnoscience-based salt hydrolysis learning. To comprehensively assess these competencies, the test items were constructed in various formats, including multiple-choice, complex multiple-choice, fill-in-the-blank, matching, and true–false questions. The use of diverse item types was intended to capture different cognitive processes, such as understanding scientific concepts, applying numerical reasoning, interpreting data, and making evidence-based decisions, detailed as follows:

Table 1. Distribution of questions

Questions Type	Sum
Multiple Choice	5
Complex Multiple Choice	5
Matching	5
Fill-in	5
True or False	5

Based on Tabel 1, the prepared question distribution, the Minimum Competency Assessment (AKM) instrument for ethnoscience-based scientific literacy and numeracy on salt hydrolysis was then developed through structured research stages. These stages were designed to ensure that each question item had a clear relationship to the competency indicators, the ethnoscience context, and the characteristics of the AKM. The research process not only focused on question formulation but also included content validation and quantitative analysis of question item quality. The research stages carried out in developing this instrument are presented in Figure 1.

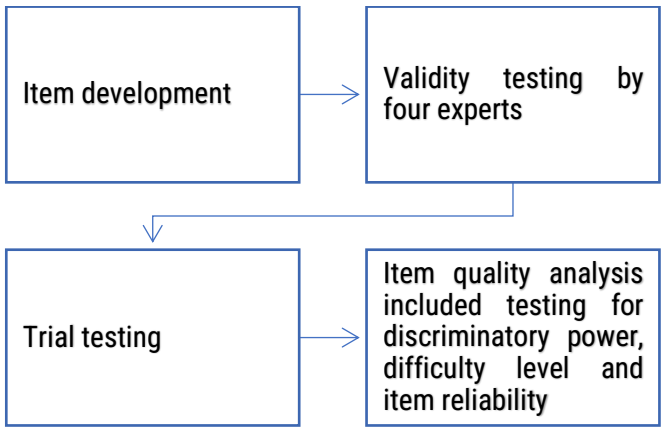

Figure 1. Stage Research

Based on the stage research, The instrument was developed using a systematic design framework integrating AKM science literacy and numeracy competencies with ethnoscience principles. The item development process consisted of four stages: (1) identification of scientific literacy and numeracy indicators based on the AKM framework, (2) selection of relevant chemistry content, specifically salt hydrolysis, (3) identification of ethnoscience contexts from Central Java local practices, and (4) item construction.

The scientific literacy indicators included explaining scientific phenomena, interpreting data and evidence scientifically, and applying scientific knowledge to real-life contexts. Meanwhile, numeracy indicators focused on quantitative reasoning, proportional reasoning, and data interpretation embedded within chemical problem situations. Ethnoscience was systematically integrated by embedding local

community practices as the primary stimulus for each item. Rather than using general contexts, each question was designed around authentic cultural activities from Central Java that conceptually relate to salt hydrolysis mechanisms. The ethnoscience integration was operationalized through contextual stimuli describing local practices such as salt production in Bledug Kuwu, the use of lerak for washing batik, the use of lime as a traditional expectorant, monosodium glutamate (MSG) in Timlo Solo, and leavening agents in Serabi Notosuman. Each context was analytically examined to identify its underlying hydrolysis principles before being transformed into AKM-type assessment items.

Table 2. Blueprint of Ethnoscience-Based AKM Instrument on Salt Hydrolysis

Explaining scientific phenomena	**Chemistry Concept (Salt Hydrolysis)**	**Ethnoscience Context**	**Focus of Explanation**
	Basic salt hydrolysis (carbonate salts)	Bledug Kuwu salt production	Explaining why certain brine samples may show basic pH due to carbonate ion hydrolysis
	Weak acid dissociation and conjugate base formation	Lime (jeruk nipis) as traditional expectorant	Explaining acid behavior and equilibrium in relation to acidity and physiological effect
	Salt from weak acid–strong base	Lerak used for washing batik	Explaining how natural surfactant solution may exhibit basic properties related to hydrolysis mechanisms
Interpreting Data and Evidence Scientifically	**Chemistry Concept (Salt Hydrolysis)**	**Ethnoscience Context**	**Type of Data Interpretation**
	pH comparison of different salt solutions	Bledug Kuwu brine samples	Interpreting experimental pH data to determine acidic, neutral, or basic salt behavior
	Hydrolysis equilibrium expression	MSG (Monosodium Glutamate) in Timlo	Analyzing pH data and identifying whether MSG solution is neutral, slightly acidic, or basic
	Relationship between concentration and pH	Lerak solution concentration	Interpreting how dilution influences hydrolysis equilibrium and solution acidity/basicity
Applying Scientific Knowledge to real-life context	**Chemistry Concept (Salt Hydrolysis)**	**Ethnoscience Context**	**Real-Life Application Focus**
	Acid–base reaction producing salt and CO_2	Serabi leavening process	Applying knowledge of acid–base reactions and salt formation to explain batter expansion
	Salt classification (strong/weak acid–base origin)	MSG in Timlo	Applying salt classification principles to predict solution properties
	Predicting pH from salt origin	Bledug Kuwu salt usage	Applying hydrolysis concept to determine suitability of salt in food processing

Instrument Validity

Before using a research instrument, a validation phase is necessary to ensure its suitability and align with the research objectives. Instrument validity is a crucial step aimed at assessing the extent to which the test items are able to measure the intended competencies. In this study, the instrument validation process was conducted by four experts with expertise in education and chemistry. These experts assessed the suitability of the test items to the indicators, ensuring each question accurately represented the competencies intended to be measured. Therefore, expert validation is a strategic step to improve the quality of the instrument before it is implemented in data collection.

The validators' assessment results will be analyzed using Aiken's formula to calculate the validity index for each item. Aiken's formula provides a quantitative value indicating the level of expert agreement on the instrument's suitability. The higher the Aiken value, the more valid the item is for use in research. The following is the Aiken's formula used:

$$V = \sum s / [n(C-1)] \text{ with } s = r - l_o$$

Description:

V = Validity coefficient
n = Number of validators
l_o = Lowest validity assessment score
C = Highest validity assessment score
r = Score given by the assessor

Discriminatory power, level of difficulty and reliability of each type of question

Discriminating power is a measure of the extent to which a test item can differentiate between competencies possessed by students, whether they have mastered the test, or have not mastered it, or those who have not mastered it. The discriminating power (DP) of each multiple-choice test item is calculated as follows:

$$DP = \frac{BA-BB}{N}$$

Where:

DP = Discriminating Power
BA = Number of participants in the upper group who answered the item correctly
BB = Number of participants in the lower group who answered the item correctly
N = Total number of test participants.

The discriminating power for complex multiple-choice, matching, fill-in-the-blank, and true-false questions is calculated using the formula:

$$DP = \frac{\text{Mean of the Upper Group - Mean of the Lower Group}}{\text{maximum score of the question}}$$

The difficulty level measurement is a calculation of the question's level of difficulty. If the question has a balanced level of difficulty, it is considered good. Questions should be neither too difficult nor too easy. The difficulty level of each multiple-choice question is calculated as follows:

$$I = \frac{B}{N}$$

Where:
I: Difficulty index for each question
B: Number of students who answered correctly
N: Total number of test participants

The difficulty level of each multiple-choice question, including complex, matching, fill-in-the-blank, and true-false statements, is calculated as follows:

$$TK = \frac{\text{Mean}}{\text{Maximum Score}}$$

$$\text{Mean} = \frac{\text{Total student scores on a question}}{\text{total number of test participants}}$$

Reliability is the condition where a test will produce the same results when administered to the same group at different times (Arifin, 2012). Reliability testing is conducted using the Cornbach Alpha formula technique with the following formula:

$$r_{11} = \left[\left(\frac{K}{K-1} \right) \right] \left[\left(1 - \frac{\sum \sigma^2 b}{\sigma^2 t} \right) \right]$$

Description:
r11 = alpha reliability coefficient
k = number of question items
∑σ^2 = number of item variants
σ^2 t = total variance

RESULT AND DISCUSSION

Instrument Validity

Based on the results of instrument validation and Aiken formula calculations, the following results as shown in table 2:

Table 2. Result Aiken Validity

Question Number	V mark	V mark table	Conclusion
1	0,91	0,75	Valid
2	0,83	0,75	Valid
3	0,83	0,75	Valid
4	0,75	0,75	Valid
5	0,91	0,75	Valid
6	0,91	0,75	Valid
7	0,83	0,75	Valid
8	0,83	0,75	Valid
9	0,83	0,75	Valid
10	0,91	0,75	Valid
11	0,75	0,75	Valid
12	0,91	0,75	Valid
13	0,83	0,75	Valid
14	0,91	0,75	Valid
15	0,83	0,75	Valid
16	0,75	0,75	Valid
17	0,83	0,75	Valid
18	0,83	0,75	Valid
19	0,75	0,75	Valid
20	0,91	0,75	Valid
21	0,75	0,75	Valid
22	0,83	0,75	Valid
23	0,83	0,75	Valid
24	0,83	0,75	Valid
25	0,83	0,75	Valid

Based on the content validity evaluation conducted using Aiken's V index, all test items satisfied

the established minimum validity criterion of V ≥ 0.75. The Aiken's V coefficients for the 25 items ranged from 0.75 to 0.91, reflecting a high degree of consistency among expert judgments regarding the relevance, clarity, and appropriateness of each item in representing the intended constructs. This range of values indicates that the experts largely agreed that the items were well-aligned with the indicators of scientific literacy and numeracy and were suitable for assessing these competencies within the ethnoscience-based salt hydrolysis topic.

The results of this analysis demonstrate that none of the items fell below the acceptable validity threshold, suggesting that all items adequately captured the essential content domains and cognitive processes targeted by the instrument. The absence of items with low Aiken's V values implies that the test items were clearly formulated, conceptually accurate, and contextually relevant, particularly in integrating ethnoscience elements with chemical concepts related to salt hydrolysis. As a result, no revisions or eliminations were required, and all 25 items were retained for subsequent stages of the research.

Furthermore, the high content validity of the instrument indicates that it is capable of measuring students' scientific literacy and numeracy skills in a comprehensive and meaningful manner. The inclusion of ethnoscience contexts within the items allows students to engage with scientific concepts through culturally relevant situations, which may enhance construct representation and reduce potential bias. This strengthens the instrument's ability to capture authentic student competencies rather than rote knowledge alone.

Overall, these findings confirm that the developed assessment instrument possesses adequate and robust content validity, fulfilling one of the essential quality criteria of educational measurement instruments. Therefore, the instrument can be considered sufficiently sound for implementation in the field data collection phase, where further quantitative analyses—such as item difficulty and discrimination indices—can be conducted to provide additional evidence of its psychometric quality and practical applicability in measuring students' scientific literacy and numeracy in the context of ethnoscience-based salt hydrolysis.

Discriminatory power, level of difficulty and reliability of each type of question

The results of the quantitative data analysis of the question items will be displayed in detail based on the question form in Table 3.

Table 3. Item Quality on Multiple Choice Questions (MCQ)

Question Numb	Type Question	TK	Category	DP	Category	Reliability	Category
1	Multiple Choice	0,67	Moderate	0,5	Good	0,758	Reliable
8		0,8	Easy	0,5	Good		
13		0,67	Moderate	0,625	Good		
18		0,76	Moderate	0,5	Good		
23		0,63	Moderate	0,5	Good		

Based on the quantitative analysis data of multiple choice questions, it can be interpreted that question number 1 has a difficulty level (TK) of 0.67 which is in the medium category, and a discriminating power (DP) of 0.5 which is in the good category. Question number 8 has a TK of 0.8 (easy), DP of 0.5 (good). Question number 13 has a TK of 0.67 (moderate), DP of 0.625 (good). Question number 18 has a TK of 0.76 (moderate), DP of 0.5 (good). Question number 23 has a TK of 0.63 (moderate), DP of 0.5 (good). The multiple choice questions have a reliability of 0.758 which indicates a reliable instrument so that the instrument is suitable for use.

Based on the results of the item analysis, it can be explained that the majority of multiple-choice questions are in the moderate difficulty category. This indicates that the questions are neither too easy nor too difficult, thus measuring student abilities more proportionally and representatively. Furthermore, all items have good discriminating power, meaning each item is able to clearly differentiate between high-ability students and low-ability students. Good discriminating power indicates that the items function

effectively in identifying differences in the level of material mastery among students.

Furthermore, the results of the reliability test indicate that the instrument as a whole has a good level of reliability, so it can be said to be consistent and stable in measuring student abilities. Therefore, the test instrument used is appropriate and reliable for use as a tool for measuring student abilities in research or learning evaluation contexts.

Table 4. Item Quality Analysis on Complex Multiple Choice Questions (CMCQ)

Question Numb	Type Question	TK	Category	DP	Category	Reliability	Category
2		0,71	Easy	0,75	Good		
6		0,55	Moderate	0,875	Good		
9	CMCQ	0,46	Moderate	0,56	Good	0,725	Reliable
14		0,56	Moderate	0,68	Good		
19		0,45	Moderate	0,5	Good		

Based on Table 4, the multiple choice questions (MCQ) can be interpreted as question number 2 having a difficulty level (TK) of 0.71, which is categorized as easy, and a discriminating power (DP) of 0.75, which is categorized as good. Question number 6 has a TK of 0.55 (moderate), DP of 0.5 (good). Question number 13 has a TK of 0.46 (moderate), DP of 0.56 (good). Question number 14 has a TK of 0.56 (moderate), DP of 0.68 (good). Question number 19 has a TK of 0.45 (moderate), DP of 0.5 (good). The complex multiple choice questions have a reliability of 0.758, indicating that the instrument is reliable and therefore suitable for use. From these results, it can be concluded that most of the complex multiple-choice questions have a difficulty level in the medium category, all questions have good discriminating power, and the instrument as a whole is reliable for use in measuring student abilities.

Table 5. Item Quality Analysis on Matching Questions

Question Number	Type Question	TK	Category	DP	Category	Reliability	Category
3		0,72	Easy	0,72	Good		
7		0,53	Moderate	0,55	Good		
10	Matching	0,63	Moderate	0,7	Good	0,789	Reliable
15		0,64	Moderate	0,7	Good		
20		0,66	Moderate	0,6	Good		

In Table 5, the matching questions can be interpreted as follows: Question 3 has a difficulty level (TK) of 0.72, which is considered easy, and a discriminating power (DP) of 0.72, which is considered good. Question 7 has a TK of 0.53 (moderate), and a DP of 0.5 (good). Question 10 has a TK of 0.63 (moderate), and a DP of 0.7 (good). Question 15 has a TK of 0.64 (moderate), and a DP of 0.7 (good). Question 20 has a TK of 0.66 (moderate), and a DP of 0.6 (good). The complex multiple-choice questions have a reliability of 0.789, indicating the instrument is reliable and suitable for use.

From these results, it can be concluded that most matching questions have a difficulty level in the moderate category, all questions have good discriminating power, and the instrument is overall reliable for use in measuring student ability.

Based on Table, the quantitative analysis data of the items in the form of fill-in questionnaires, it can be interpreted that question number 4 has a difficulty level (TK) of 0.64 which is in the medium category, and a discriminating power (DP) of 0.52 which is in the good category. Question number 11 has a TK of 0.6 (moderate), DP of 0.525 (good). Question number 16 has a TK of 0.58 (moderate), DP of 0.47 (good). Question number 21 has a TK of 0.63 (moderate), DP of 0.42 (good). Question number 22 has a TK of 0.76 (easy), DP of 0.65 (good). The fill-in questionnaire has a reliability of 0.71, indicating that the instrument is reliable and therefore suitable for use

Table 6. Item Quality Analysis on Fill-in Questions

Question Number	Type Question	TK	Category	DP	Category	Reliability	Category
4	Fill-in	0,64	Moderate	0,525	Good	0,71	Reliable
11		0,6	Moderate	0,525	Good		
16		0,58	Moderate	0,47	Good		
21		0,63	Moderate	0,42	Good		
22		0,76	Easy	0,65	Good		

Table 7. Item Quality Analysis on True or False Questions

Question Number	Type Question	TK	Category	DP	Category	Reliability	Category
5	True or False	0,76	Easy	0,67	Good	0,814	Reliable
12		0,68	Moderate	0,67	Good		
17		0,68	Moderate	0,62	Good		
22		0,76	Easy	0,67	Good		
24		0,89	Easy	0,8	Good		

Based on Table 7, the true-false items can be interpreted as having a difficulty level (TK) of 0.76, which is categorized as easy, and a discriminating power (DP) of 0.67, which is categorized as good. Question 12 has a TK of 0.68 (moderate), and a DP of 0.67 (good). Question 17 has a TK of 0.68 (moderate), and a DP of 0.62 (good). Question 22 has a TK of 0.76 (easy), and a DP of 0.67 (good).

Question 24 has a TK of 0.89 (easy), and a DP of 0.8 (good). The fill-in-the-blank items have a reliability of 0.814, indicating the instrument is reliable and suitable for use. From these results, it can be concluded that most fill-in-the-blank items have a difficulty level in the easy category, all items have good discriminating power, and the instrument is overall reliable for use in measuring student ability.

Based on the analysis of the quality of multiple-choice test items, it was found that most of the test items were moderately difficult, with the difficulty index values falling between the limits of easy and difficult. This indicates that the multiple-choice test items were able to measure students' abilities proportionally. Furthermore, all multiple-choice test items had good discriminatory power, indicating that they were able to clearly differentiate between students with high and low abilities. A reliability value of 0.758 indicates that the multiple-choice test has good internal consistency, making it suitable for use as a measure of student ability. This finding aligns with Arikunto's (2015) opinion that good test items generally have moderate difficulty and adequate discriminatory power.

In the complex multiple-choice (MCC) format, analysis results indicate that most items are in the moderate difficulty category, although one item is in the easy category. The discriminatory power of all MCC items is in the good category, indicating that they are effective in differentiating student ability levels. This indicates that, despite the higher complexity of MCC questions, they still function effectively as an evaluation tool. A reliability value of 0.758 indicates that the MCC instrument is reliable and consistent in measuring student ability. According to Sudijono (2011), good discriminatory power is an important indicator that an item is worthy of being retained in a test instrument.

Analysis results for the matching questions indicate that most items are in the moderate difficulty category, with one item in the easy category. All items have good discriminatory power, indicating that matching questions are able to effectively measure differences in student ability. A reliability value of 0.789 indicates that the matching question format has a good level of reliability. This shows that the matching question format can be used optimally to measure students' understanding of interrelated concepts.

Furthermore, in the form of fill-in-the-blank questions, the analysis results show that most of the items have a medium level of difficulty, with one item in the easy category. The discriminatory power of all fill-in-the-blank questions is in the good category, indicating that fill-in-the-blank questions are able to

reveal students' abilities in more depth, especially in understanding and remembering concepts without the help of answer choices. A reliability value of 0.71 indicates that the fill-in-the-blank instrument is classified as reliable. This is in line with the opinion of Arikunto (2015) who stated that fill-in-the-blank questions with good discriminatory power can provide a more authentic picture of students' abilities.

For the true-false statement questions, the analysis results indicate that the difficulty level tends to be moderate to easy. All items have good discriminating power, indicating that despite the relatively simple format, the questions are still able to effectively differentiate student abilities. A reliability value of 0.814 indicates that the true-false instrument has very good consistency. Therefore, the true-false question format remains suitable for use as part of an evaluation instrument if it is formulated with clear statements and does not allow for multiple interpretations.

Overall, the analysis results of the difficulty level, discriminating power, and reliability indicate that the test instrument used in this study is of good quality. The predominance of items with moderate difficulty, good discriminating power, and high reliability values indicate that the test instrument is suitable for measuring student abilities validly and reliably. These findings confirm that the instrument meets the criteria for a good test as outlined in educational evaluation theory. Despite the promising findings regarding item difficulty, discriminating power, and reliability, this study has several limitations that should be considered when interpreting the results. First, the instrument was developed and tested within a specific regional context, namely Central Java, using ethnoscience practices such as Bledug Kuwu salt production, lerak for batik washing, lime as a traditional remedy, MSG in Timlo Solo, and Serabi Notosuman preparation. Therefore, the contextual relevance of the items may not be directly transferable to students from different cultural backgrounds. Second, the study focused primarily on classical test theory analysis to evaluate item quality. Although the results indicate good psychometric properties, more advanced modeling approaches such as Rasch analysis could provide deeper insights into item functioning and measurement invariance. Third, this research emphasizes instrument development and psychometric evaluation, but does not yet examine the long-term impact of ethnoscience-based assessment on students' scientific literacy development. Future research is needed to explore how culturally grounded assessment influences conceptual understanding and higher-order thinking skills over time. These limitations suggest that while the instrument demonstrates good technical quality within the studied context, further validation and broader implementation studies are required to strengthen its generalizability and educational impact.

CONCLUSION

The assessment instrument developed in this study demonstrated satisfactory quality for measuring students' scientific literacy and numeracy skills. The 25 test items, constructed in various formats, successfully represented the targeted competencies and underwent a rigorous validation process. Expert judgment analysis using Aiken's validity index confirmed that all items met the required content validity criteria. Furthermore, the results of the empirical item analysis showed that the test items possessed appropriate levels of difficulty, good discriminatory power, and high reliability. These findings indicate that the instrument is both valid and reliable, making it suitable for use in assessing students' scientific literacy and numeracy within the ethnoscience-based salt hydrolysis context. Consequently, the developed instrument can be effectively utilized in educational evaluation and further research to support the measurement and improvement of students' competencies.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to all parties who contributed to the completion of this study. Special appreciation is extended to the expert validators for their valuable time, insights, and constructive feedback in evaluating the assessment instrument. The authors also thank the teachers and students who participated in the field trial for their cooperation and willingness to support this research. Furthermore, gratitude is conveyed to colleagues and academic peers for their discussions and suggestions that helped improve the quality of this work. This study was conducted with the support of the affiliated institution, which provided academic and administrative assistance throughout the

research process.

REFERENCES

Asrijanty. (2020). *AKM dan Implikasinya pada Pembelajaran*. Pusat Asesmen Dan Pembelajaran Kemendikbud

Dewi, C. A., Khery, Y., & Erna, M. (2019). An ethnoscience study in chemistry learning to develop scientific literacy. *Jurnal Pendidikan IPA Indonesia*, *8*(2), 279–287. https://doi.org/10.15294/jpii.v8i2.19261

Dewi Purwati, P., Faiz, A., Widiyatmoko, A., & Maryatul, S. (2021). Asesmen Kompentensi Minimum (AKM) kelas jenjang sekolah dasar sarana pemacu peningkatan literasi peserta didik. *Jurnal Kajian Pendidikan Umum*, 13–24

OECD. (2015a). *PISA Result in Focus*. OECD Publisher

OECD. (2015b). *Programme For International Student Assessment (PISA) Result From PISA 2015*. OECD Publisher.

OECD. (2018). *Programme For International Student Assessment (PISA) Results From PISA 2018*. OECD =Publisher

OECD. (2023). *Programme For International Student Assessment (PISA) Results From PISA 2023*. OECD Publisher

Sudarmin. (2014). *Pendidikan Karakter, Etnosains dan Kearifan Lokal (Konsep dan Penerapannya dalam Penelitian dan Pembelajaran Sains)* (Vol. 1). Pengetahuan Alam Universitas Negeri Semarang.

Sumarni, W., Sudarmin, Wiyanto, Rusilowati, A., & Susilaningsih, E. (2017). Chemical literacy of teaching candidates studying the integrated food chemistry ethnosciences course. *Journal of Turkish Science Education*, *14*(3), 60–72. https://doi.org/10.12973/tused.10204a

Ulyah, S. M., Sediono, S., Ana, E., Sholihah, N., & Niswatin, K. (2021). Improving the Competency of High School Teachers in Understanding and Designing Questions Based on Minimum Competency Assessment in Babat Lamongan District. *MUST: Journal of Mathematics Education, Science and Technology*, *6*(1), 55. https://doi.org/10.30651/must.v6i1.7773

Yamtinah, S., Masykuri, M., & Syahidul Shidiq, A. (2017). *An Analysis of Students' Science Process Skills in Hydrolysis Subject Matter Using Testlet Instrument*