MJP Journal of Education and Teaching Learning
Vol. 4, No. 1 (2026): 112 – 121
https://jurnalilmiah.co.id/index.php/MJPJETL

Artificial Intelligence-Based Scientific Literacy Assessment Integrated with Papuan Ethnochemistry: Content Validity and Readability in Stoichiometry

[1] Cartika Candra Ledoh, [2] Suriyah Satar, [3] Hanida Listiani, [4] Nurbaya Nurbaya, [5] Mariana Nensi, [6] Mutiara Dama Agih

[1,5,6] Chemistry Education Study Program, Faculty of Teacher Training and Education, Universitas Cenderawasih, Jayapura, Indonesia
[2,3,4] Biology Education Study Program, Faculty of Teacher Training and Education, Universitas Cenderawasih, Jayapura, Indonesia

***Corresponding Author:** cartikacandraledoh1@gmail.com

Article Info

Keywords:

First keyword;
Second keyword;
Third keyword;
Fourth keyword;
Fifth keyword.

ABSTRACT

This study aimed to develop a scientific literacy assessment instrument for stoichiometry integrated with Papuan ethnochemistry and supported by Artificial Intelligence (AI) for preliminary scoring and feedback provision. The study employed a Research and Development (R&D) method adapted from the ADDIE model and was limited to the Development stage. The developed product consisted of 15 open-ended items equipped with Papuan ethnochemistry stimuli, scientific literacy indicators, answer keys, an analytic scoring rubric, and a structured prompt for Google Gemini. Content validation involved 10 validators, comprising seven chemistry education lecturers and three high school chemistry teachers. Content validity was analyzed using Aiken's V index, with an operational cut-off value of 0.73. A limited trial involved 15 Grade 11 students representing high, moderate, and low academic ability to obtain data on readability and initial responses. The results showed that the Aiken's V values for all items ranged from 0.85 to 0.97, with a mean value of 0.91, indicating that all 15 items met the content validity criteria. The limited trial showed that most students were able to understand the instructions, stimuli, questions, and Papuan ethnochemistry contexts. Students also responded positively to the use of cultural contexts and AI-based feedback. The identified difficulties mainly involved transforming narrative information into calculation procedures, determining relevant data, converting mass to moles, applying coefficient ratios, and interpreting calculation results. Therefore, the developed instrument demonstrated preliminary evidence of good content validity and readability and has the potential to be used as a contextual scientific literacy assessment for stoichiometry. However, the accuracy and consistency of AI scoring compared with human raters, as well as the effectiveness of the instrument in improving students' scientific literacy, require further empirical investigation.

Article Information

Keywords:
First keyword;
Second keyword;
Third keyword;
Fourth keyword;
Fifth keyword.

ABSTRACT

This study aimed to develop a scientific literacy assessment instrument for stoichiometry integrated with Papuan ethnochemistry and supported *by Artificial* Intelligence (AI) for preliminary scoring and feedback provision. The study employed *a Research and* Development (R&D) method by adapting the ADDIE model and was *limited to* the Development stage. The developed product comprised 15 open-ended items equipped with Papuan ethnochemistry stimuli, scientific literacy indicators, answer *keys,* an analytic scoring rubric, and a structured prompt for Google Gemini. Content validation involved 10 validators, consisting of seven chemistry education lecturers and three high *school* chemistry teachers. Content validity was analyzed using Aiken's V index with an operational cut-off value of 0.73. A limited trial involved 15 Grade 11 students representing high, moderate, and low academic ability to obtain data on readability and initial responses. The results showed that the Aiken's V values for all items ranged from 0.85 to 0.97, with a mean of 0.91; thus, all 15 items met the content validity criteria. The limited trial showed that most students were able to understand the instructions, stimuli, questions, and Papuan ethnochemistry contexts. Students also responded positively to the use of cultural contexts and AI-based feedback. The difficulties identified mainly involved transforming narrative information into calculation procedures, determining relevant data, converting mass to moles, applying coefficient ratios, and interpreting calculation results. Therefore, the developed instrument demonstrated preliminary evidence of good content validity and readability and has the potential to be used as a contextual scientific literacy assessment for stoichiometry. However, the accuracy and consistency of AI scoring relative to human raters, as well as the effectiveness of the instrument in improving students' scientific literacy, still require empirical testing in further research.

Article History

Received : 24/07/2026
Revised : 10/08/2026
Accepted : 21/08/2026

✉ ***Corresponding Author:*** (1) Cartika Candra Ledoh, (2) Chemistry Education Study Program, (3) Universitas Cenderawasih, , (4) Jayapura, Indonesia, (5) Email: cartikacandraledoh1@gmail.com

1. Introduction

Scientific literacy is an essential competency in 21st-century education, as it enables students to use scientific knowledge to explain phenomena, evaluate information, interpret evidence, and make responsible decisions. The results of the Programme for International Student Assessment (PISA) 2022 indicate that Indonesian students' science performance remains below the OECD average [1]. This condition suggests that science teaching and assessment have not yet fully prepared students to apply scientific concepts to real-world problems. Fuadi et al. [2] emphasize that low scientific literacy is influenced not only by students' conceptual knowledge but also by limited opportunities to engage with contextual problems, use evidence, and communicate scientific reasoning. Similar findings have been reported across different educational levels in Indonesia, from junior to senior high school, consistently indicating low levels of scientific literacy across various science topics [16]–[19].

Assessment plays a strategic role in strengthening scientific literacy because the format of assessment determines the types of abilities that are practiced and valued in learning. Assessments that predominantly emphasize memorization and routine calculations tend to provide limited information about students' ability to apply chemical knowledge. The chemistry literacy taxonomy proposed by Shwartz et al. [3] indicates that chemical literacy should encompass conceptual understanding, the use of chemical language, the application of knowledge to everyday life, and the ability to critically evaluate information. Therefore, assessment instruments in chemistry education need to be designed to elicit scientific explanations, data interpretation, evidence evaluation, and decision-making. The PISA assessment framework further emphasizes that scientific literacy includes the competencies of explaining phenomena scientifically, evaluating and designing scientific investigations, and interpreting data and evidence scientifically, which serve as references for developing the indicators used in this instrument [20].

Stoichiometry is a fundamental area of chemistry that connects macroscopic, submicroscopic, and symbolic representations. Students are required not only to understand the concepts of moles, molar mass, chemical equations, limiting reactants, and reaction yields but also to select relevant information and establish quantitative relationships among substances. Learning difficulties in basic chemical laws and stoichiometry often arise when students memorize algorithms without understanding the underlying concepts and problem contexts [5]. Consequently, students may be able to solve problems with patterns similar to those presented in examples but experience difficulties when information is presented through narratives, data, or everyday phenomena. Diagnostic studies have also identified various misconceptions in stoichiometry, particularly concerning limiting reactants, the relationship between mass and moles, and the interpretation of chemical equations [21]–[25]. These findings indicate that difficulties in stoichiometry are conceptual in nature and cannot be adequately addressed through procedural practice alone.

One approach to making assessment more contextual is ethnochemistry. Ethnochemistry connects chemical concepts with the knowledge, materials, processes, and cultural practices of communities. The identification of local cultural practices indicates that various community activities can be explained through concepts involving substances, changes in matter, chemical reactions, composition, and quantitative relationships [6]. Ethnochemistry-based learning can increase the relevance of learning because students encounter scientific concepts through cultural experiences that are familiar to them [7]. In the Papuan context, sago processing, traditional wood burning, fish smoking, and the use of slaked lime (kapur sirih) have the potential to serve as stimuli for stoichiometry assessment. Various ethnochemistry studies in In*donesia* have also demonstrated the potential of local wisdom from different regions, including traditional food processing and community cultural practices, as contextual sources for chemistry learning [26]–[31].

The use of local culture in assessment must, however, be implemented carefully. Cultural contexts should not merely serve as narrative backgrounds that can be removed without affecting the solution to the problem. Instead, the context should provide information, data, and problems that students genuinely need to analyze. Morales-Doyle et al. [13] explain that the use of sociocultural contexts in science education should maintain scientific accuracy while appropriately representing community practices. Therefore, the integration of ethnochemistry into assessment requires validation not only of the chemical concepts but also of the clarity and appropriateness of the cultural representation.

The development of Artificial Intelligence (AI) has created new opportunities for assessment practices, particularly for preliminary scoring and feedback provision. AI can process students' responses efficiently and generate more specific comments when provided with answer keys, indicators, and clear scoring rubrics. Dwivedi et al. [9] position AI as a technology that can support decision-making but requires governance, transparency, and human oversight. In educational assessment, AI has the potential to support adaptive feedback and improve evaluation efficiency [10]. However, AI-generated scores should not be treated as final decisions before their consistency and agreement with human raters have been examined. Several systematic reviews have reported that generative AI-based feedback can improve students' motivation and the quality of their revisions and can demonstrate relatively high agreement with human assessment across various educational contexts [32]–[36]. Nevertheless, other studies emphasize the need for ethical frameworks and human oversight to prevent excessive dependence on automated systems [37].

Several studies have developed scientific literacy assessment instruments based on social and cultural contexts [8], identified the potential of ethnochemistry for learning [6], and examined the use of AI in assessment [9], [10]. Nevertheless, the integration of these three elements into an open-ended stoichiometry assessment using Papuan cultural contexts remains limited. This study addresses this gap by developing 15 open-ended assessment items designed to measure scientific literacy competencies, use Papuan ethnochemistry contexts as stimuli, and incorporate an analytic scoring rubric and a structured AI prompt for generating preliminary scores and feedback.

This study aims to (1) describe the need for developing a scientific literacy assessment instrument for stoichiometry; (2) develop the characteristics of an AI-supported scientific literacy assessment product integrated with Papuan ethnochemistry; (3) analyze the content validity of the instrument based on expert judgments; and (4) describe the readability and initial student responses during a limited trial. The findings are expected to provide a foundation for further empirical testing and broader implementation of the instrument.

2. Research Method

This study employed a Research and Development (R&D) method adapted from the ADDIE model [14]. The ADDIE model consists of five stages: Analysis, Design, Development, Implementation, and Evaluation. The study was limited to the Development stage because its primary focus was to produce an initial assessment product, obtain evidence of content validity, and describe the readability of the instrument. Large-scale Implementation and Evaluation stages were not conducted in this study.

The study was conducted at SMA YPPK Taruna Dharma Jayapura. The data source for the needs analysis stage was a Grade 11 chemistry teacher. Content validation involved 10 validators, comprising seven chemistry education lecturers from different universities and three high school chemistry teachers with experience teaching stoichiometry. A limited trial involved 15 Grade 11 students representing high, moderate, and low academic ability. The students were purposively selected to ensure that the readability assessment covered a range of academic abilities.

The choice of this development procedure is consistent with the development of educational assessment instruments, which generally involves needs analysis, design, development, validation, and initial testing before large-scale implementation [14]. In developing scientific literacy assessment instruments, alignment among the construct, indicators, content, language, and student characteristics is an important aspect that should be considered before the instrument is implemented more broadly [8], [15].

2.1 Validator Characteristics and Qualifications

The validators were selected based on the relevance of their backgrounds to the aspects being assessed. Seven validators specialized in chemistry education, while three were high school chemistry teachers with experience teaching stoichiometry. The validation aspects included alignment of the content with the learning outcomes, the appropriateness of the scientific literacy indicators, the accuracy of the integration of Papuan ethnochemistry, item construction, language clarity, the appropriateness of the scoring rubric, and the suitability of the AI system for evaluating students' responses.

Based on the available research documentation, the explicitly documented qualifications were a chemistry education background for the seven lecturers and experience teaching stoichiometry for the three high school chemistry teachers. Individual information regarding years of experience, specific expertise in scientific literacy assessment, or research experience in stoichiometry for each validator was not documented in the current research manuscript. Therefore, these details were not estimated in this study and should be added from the validators' biographies or curricula vitae, if available.

More specifically, the relevant competencies for instrument development included an understanding of stoichiometric concepts, scientific literacy assessment, the construction of open-ended items, the development of scoring rubrics, and culturally contextualized learning. The involvement of subject-matter experts and educational practitioners was necessary because content validity requires expert judgment regarding the alignment between the substance of the items and the construct being measured [11].

Table 1. Characteristics of the Validators

Validator Group	Number	Relevant Qualification/Expertise
Chemistry Education Lecturers	7	Chemistry education background; affiliated with different universities
High School Chemistry Teachers	3	High school chemistry teachers with experience teaching stoichiometry
Total	10	Content validation panel for the assessment instrument

2.2 Content Validation Process

The validators evaluated each item in terms of content, scientific literacy indicators, the Papuan ethnochemistry context, item construction, language, answer keys, scoring rubrics, and the AI-based scoring structure. The assessment used four relevance categories: Not Relevant (NR), scored 1; Less Relevant (LR), scored 2; Quite Relevant (QR), scored 3; and Relevant (R), scored 4.

Validation was conducted in two rounds. The first round was conducted to obtain evaluations of the initial draft and suggestions for improvement. Based on the results, the instrument was revised in terms of language, stimulus clarity, symbols and units, ethnochemistry contexts, simplification of numerical calculations, and rubric descriptors. The revised draft was subsequently assessed in the second round to confirm the improvements made. The validation process was conducted from March to May 2026.

2.3 Aiken's V Analysis and Decision Cutoff

Content validity data were analyzed using the Aiken's V index [11] according to the following formula:

$V = \Sigma s/[n(c-1)]$,

where $s = r - l_0$; r is the score assigned by each validator, l_0 is the lowest score on the rating scale, c is the highest score on the rating scale, and n is the number of validators. Aiken's V values range from 0 to 1, with values closer to 1 indicating higher content validity [11]. In this study, $V \geq 0.73$ was used as the operational cut-off criterion for retaining an item, whereas $V < 0.73$ was used as the basis for revision or consideration of item elimination. The value of 0.73 was not treated as a universal Aiken's V threshold. Instead, it was defined as an operational decision criterion for this study because the interpretation of Aiken's V depends on the number of validators, the number of response categories, and the purpose of the measurement [11].

Table 2. Content Validity Decision Criteria

No.	Aiken's V Interval	Decision
1	V ≥ 0,73	Valid/Accepted
2	V < 0,73	Revised or eliminated

2.4 Instrument and Scoring Rubric

The research data consisted of quantitative and qualitative data. Quantitative data comprised expert validation scores, whereas qualitative data consisted of teacher interview results, validator comments, observations during the limited trial, and student responses. The research instruments included a needs-analysis interview guide, an expert validation sheet, a limited-trial observation sheet, and an open-ended interview guide.

An analytic scoring rubric was used to assess four aspects of students' responses: conceptual accuracy, completeness of the solution process, use of data or evidence, and the connection between the response and the context. Each aspect was scored from 0 to 3, resulting in a theoretical maximum score of 12 points per item. The rubric was used both for human assessment and as the basis for the AI scoring instructions.

This approach is consistent with scientific literacy assessment, which requires students to explain phenomena, evaluate information, and use evidence [3], [15], [20]. To ensure scoring reproducibility, four performance levels were used as follows.

Table 3. Analytic Rubric for Scoring Students' Responses

Aspect	Score			
	3	2	1	0
Conceptual Accuracy	The stoichiometric concept is correct, and all relationships among quantities are accurate.	The main concept is correct, but there is a minor error.	Demonstrates partial understanding of the concept but contains a substantive error.	Does not demonstrate a relevant concept.
Completeness of the Solution Process	The solution steps are complete, logical, and sequential.	Most steps are correct, but one step is incomplete.	Some solution steps are correct, but the relationships between steps are unclear.	Does not demonstrate a relevant solution process.
Data/Evidence	All relevant data are selected and used appropriately.	The main data are used, but there is one omission or error.	Only some of the data are used appropriately.	Does not use relevant data/evidence.
Connection to Context	The result is appropriately connected to the given ethnochemical phenomenon.	A connection to the context is present but incomplete.	The connection to the context is general or partially inappropriate.	Does not connect the response to the context.

Application of the rubric: A response receives a score of 3 for conceptual accuracy when the student identifies the appropriate stoichiometric concept, correctly applies the relationship involving moles, and demonstrates no conceptual errors. A score of 1 is assigned when the student demonstrates partial understanding of the relevant concept but makes a substantive error that affects the solution. Anonymized examples of students' responses would be more appropriately presented in an appendix because the raw student responses were not available in the research source used for this revision.

The use of an explicit scoring rubric is important because variations in response characteristics may influence automated scoring outcomes. Research on machine-learning scoring in science assessment indicates that construct characteristics and response formats need to be considered when developing automated scoring systems [36], [38], [39].

2.5 Implementation of AI-Based Assessment

The structure of the Artificial Intelligence (AI)-based assessment prompt consisted of R*ole and Objective, Learning Context, Question and Answer Key, Scientific Literacy Indicator, Scoring Rubric, Evaluation Instructions, Feedback Structure,* and Final *Comment.* These components were used to guide the system in evaluating students' responses based on the ethnochemistry context, scientific literacy indicators, answer key, and predetermined scoring rubric.

In this study, Google Gemini was used as an AI system to support the preliminary scoring and provision of feedback on students' open-ended responses. Each student response was processed individually using the same structured prompt. The prompt instructed Gemini to evaluate the response based on the evidence explicitly provided by the student and not to assume solution steps that were not written in the response.

Gemini received the following inputs: (1) the question text; (2) the ethnochemistry context presented in the stimulus; (3) the scientific literacy indicator being assessed; (4) the answer key or elements of an ideal response; (5) the analytic scoring rubric; and (6) the student's response. Based on these inputs, the system was instructed to evaluate conceptual accuracy, completeness of the solution process, use of data or evidence, and connection to the context.

The system output was structured to include scores for each rubric aspect, the total score, strengths of the response, weaknesses of the response, and suggestions for improvement. This output structure was intended to assist teachers in obtaining preliminary information about the quality of students' responses while providing feedback that students could understand.

The use of Gemini as an assistive tool rather than a replacement for teachers was based on the need for transparency, human oversight, and evaluation of potential bias in AI-based assessment [9], [10], [36], [37], [40]. Furthermore, automated scoring of scientific responses needs to be evaluated against human scoring because construct characteristics and response formats may influence automated scoring outcomes [36], [38], [39]. The structure of the AI-based assessment prompt is presented in Table 4.

Table 4. Specifications of the Gemini-Based Assessment System

Component	Specification
AI System	Google Gemini
Function	Preliminary scoring and provision of feedback on open-ended responses
Input	Questions, stimuli, scientific literacy indicators, answer keys, analytic scoring rubric, and student responses
Prompt	Structured prompt with Role a*nd Objective, Learning Context, Question and Answer Key, Scientific Literacy Indicator, Scoring Rubric, Evaluation Instructions, Feedback Structure, and* Final *Comment*
Processing Unit	One student response per scoring process
Scoring Basis	Analytic scoring rubric predetermined by the researchers
Score per Aspect	0-3
Assessed Aspects	Conceptual accuracy, process completeness, use of data/evidence, and connection to context
Total Score	0-12
Qualitative Information	Strengths, weaknesses, and suggestions for improvement
Human Oversight	AI results are positioned as preliminary scores and require human verification

2.6 Selection of Subjects for the Limited Trial

The limited trial involved 15 Grade 11 students representing high, moderate, and low academic ability levels. Participants were purposively selected to obtain variation in students' abilities during the readability assessment.

In the current research manuscript, the classification consisted of three groups, with five students in each group. The basis for classification and the scores used should follow the original research records. If the study used chemistry report-card scores from the previous semester with the classification presented in the manuscript, the following criteria can be applied.

Table 5. Criteria for Selecting Students for the Limited Trial

Category	Chemistry Score Criterion	Number	Proportion
High	≥ 85	5	33,3%
Moderate	75-84	5	33,3%
Low	< 75	5	33,3%
Total		15	100%

The balanced composition of participants was intended to ensure that the readability of the instrument could be examined across different levels of academic ability. Students with high academic ability were included to assess readability among students who could relatively quickly understand the relationships among the stimulus, concepts, and calculation procedures. The moderate group represented students with average academic ability, whereas the low-ability group was included to identify parts of the instrument that might cause comprehension **difficulties.**

2.7 Guidelines for Collecting Readability Data

Readability data were obtained through observations during the assessment and open-ended interviews conducted after students completed the instrument. The observed aspects included the clarity of instructions, comprehensibility of stimuli and questions, understanding of chemical and cultural terms, presentation, ease of access via mobile phones, and students' understanding of AI-generated scores and feedback.

The readability instrument was developed based on the principle that an assessment should be understandable to students without imposing additional linguistic demands unrelated to the construct being measured. This aspect is particularly important because difficulties in chemistry may arise when students encounter complex information and representations, including in stoichiometry [21]–[24].

Table 6. Indicators of Instrument Readability

Indicator	Observation Focus
Instructions	Students can follow the instructions without repeated additional explanations.
Stimulus	Students can identify the phenomena, information, and important data.
Questions	Students understand what is being asked before beginning the calculation.
Chemical Terms	Terms, symbols, units, and notation can be understood.
Cultural Terms	The ethnochemistry context is recognized and does not cause misinterpretation.
Presentation	The arrangement of text, images, tables, and answer-input fields is easy to read.
Access	The instrument can be accessed and completed using a mobile phone.
AI Feedback	Students understand the scores, strengths, weaknesses, and suggestions for improvement.

The open-ended interview guide focused on the following questions: (1) which part of the instrument was easiest to understand; (2) which part was confusing; (3) whether the cultural context helped students understand the problem; (4) whether the available data were sufficient to begin the calculation; (5) whether the chemical terms were easy to understand; (6) whether the AI-generated scores and feedback were understandable; and (7) which aspects of the instrument needed improvement.

3. Results and Discussion

The needs analysis conducted through interviews with a Grade 11 chemistry teacher indicated that stoichiometry assessment was still primarily oriented toward conceptual mastery and calculation skills. Assessment formats were dominated by multiple-choice questions and simple open-ended questions, whereas students' abilities to explain scientific phenomena, interpret data, use evidence, and apply chemical concepts to real-world situations had not been specifically assessed.

These findings indicate a need to develop an assessment that goes beyond numerical answers and requires students to explain the relationship between chemical concepts and the phenomena presented in the problems. This need is consistent with the characteristics of scientific literacy, which require students to use knowledge to explain phenomena, evaluate information, and use evidence [3], [8], [15], [20]. The findings are also consistent with previous studies indicating that Indonesian students' scientific literacy requires further strengthening [16]–[19].

This need is particularly important in stoichiometry because students are required to connect chemical representations with quantitative relationships. Previous studies have reported difficulties and misconceptions concerning the relationships among mass, moles, chemical equations, and stoichiometric problem solving [21]–[24]. Therefore, assessment should integrate conceptual understanding, data interpretation, procedural skills, and scientific reasoning.

The development of the instrument was also based on the need to integrate local Papuan culture. Teachers had previously connected chemistry learning with sago, traditional firewood burning, marine-product processing, and other traditional activities; however, these contexts had not yet been systematically incorporated into assessment instruments.

These findings support the use of ethnochemistry as an approach to connecting chemical concepts with community practices and knowledge [6], [7], [26], [28], [29]. Studies of local wisdom have also demonstrated that food-processing activities and cultural practices can be developed as contextual resources for science and chemistry learning [27], [30], [31].

Thus, the development needs identified in this study encompass three main aspects: strengthening scientific literacy assessment, integrating Papuan ethnochemistry contexts, and utilizing technology to support scoring and feedback processes.

3.1 Characteristics of the Developed Product

The developed product consisted of 15 open-ended scientific literacy assessment items on stoichiometry. Each item included scientific literacy indicators, Papuan ethnochemistry stimuli, questions, answer keys, scoring rubrics, and an AI feedback format.

The targeted competencies included the ability to explain phenomena scientifically, evaluate or design scientific investigations, and interpret scientific data and evidence [3], [15], [20].

The contexts used in the instrument included sago processing, traditional wood burning, fish smoking, and the use of slaked lime (k*apur* sirih). These contexts were selected because they are closely related to students' immediate surroundings and can be connected to stoichiometric concepts. The use of these local contexts is consistent with ethnochemistry studies demonstrating the potential of cultural practices as contextual resources for chemistry learning [6], [7], [26], [28]–[31].

An open-ended format was employed to allow students to demonstrate their reasoning, solution processes, use of data, and connections between their responses and the given phenomena. Thus, the instrument can provide richer information about students' reasoning than selected-response formats alone [3], [8], [15].

3.2 Content Validity of the Instrument

Content validation involved 10 validators, consisting of seven chemistry education lecturers and three high school chemistry teachers. The three teachers had experience teaching stoichiometry. The assessment covered content relevance, scientific literacy indicators, integration of Papuan ethnochemistry, item construction, language, scoring rubrics, and the AI-based assessment system.

The assessment used four categories: Not Relevant (1), Less Relevant (2), Quite Relevant (3), and Relevant (4). The Aiken's V index was used to analyze the content relevance of each item [11].

The analysis showed that the Aiken's V values for all items ranged from 0.85 to 0.97, with a mean value of 0.91. The lowest value was obtained for Item 1 (0.85), whereas the highest was obtained for Item 11 (0.97). All values exceeded the operational cut-off value of 0.73, indicating that all 15 items met the predetermined content validity criterion.

Table 7. Aiken's V Analysis of the Scientific Literacy Assessment Instrument

Item	Aiken's V	Criterion	Conclusion
1	0,85	≥ 0,73	Valid
2	0,87	≥ 0,73	Valid
3	0,89	≥ 0,73	Valid
4	0,91	≥ 0,73	Valid
5	0,88	≥ 0,73	Valid
6	0,95	≥ 0,73	Valid
7	0,92	≥ 0,73	Valid
8	0,90	≥ 0,73	Valid
9	0,89	≥ 0,73	Valid
10	0,96	≥ 0,73	Valid
11	0,97	≥ 0,73	Valid
12	0,93	≥ 0,73	Valid
13	0,91	≥ 0,73	Valid
14	0,90	≥ 0,73	Valid
15	0,95	≥ 0,73	Valid

Item	Aiken's V	Criterion	Conclusion
Mean	0,91	≥ 0,73	High content validity

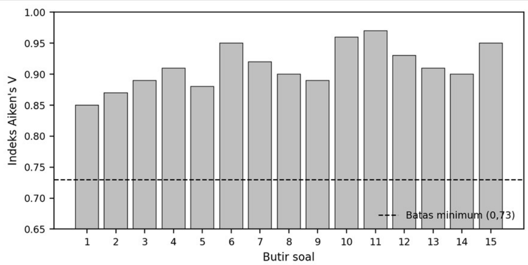

Figure 1. Distribution of Aiken's V Values for Each Assessment Item

The distribution indicates that none of the items approached the operational cut-off value. The lowest value, observed for Item 1, was 0.85, which was 0.12 points above the decision criterion. Most items had Aiken's V index values of 0.90 or higher.

These findings provide preliminary evidence that the expert panel considered the content of the instrument to be aligned with the intended construct. This finding supports previous research on scientific literacy assessment development, which emphasizes the importance of alignment among indicators, content, language, and student characteristics [8], [15].

3.3 Revisions Based on Validator Feedback

Although all items met the content validity criterion, the validators provided several suggestions for improvement. These included simplifying the language, adjusting scientific terminology, incorporating more familiar Papuan cultural contexts, improving illustrations, simplifying numerical values in stoichiometric calculations, standardizing symbols and units, refining the analytic scoring rubric, and strengthening the AI feedback format.

These findings demonstrate that quantitative evidence of content validity does not eliminate the need for qualitative revision. Simplifying the language was necessary to ensure that difficulties in understanding sentence structures did not become sources of difficulty unrelated to scientific literacy. Similarly, simplifying numerical data was necessary to ensure that the assessment emphasized reasoning and conceptual application rather than merely mathematical operations. This consideration is relevant to previous findings concerning students' difficulties with stoichiometry [21]–[24].

Regarding the ethnochemistry context, the validators' comments also indicate that the mere inclusion of cultural terminology is insufficient. The context must be scientifically accurate, relevant to the chemistry problem, and genuinely necessary for students to solve the problem [13].

3.4 Integration of Ethnochemistry into the Instrument

The integration of ethnochemistry was one of the main characteristics of the developed instrument. Sago processing, traditional wood burning, fish smoking, and the use of slaked lime (k*apur sirih)* were selected because these practices are connected to chemical concepts and familiar to students.

The use of these contexts is consistent with previous studies showing that cultural practices can be developed as resources for chemistry learning [6], [26], [28], [29]. The ethnochemistry approach can also strengthen the relationship between scientific concepts and students' social and cultural experiences [7].

In the Papuan context, the use of stimuli familiar to students has additional relevance. Research on the scientific literacy profiles of Papuan students highlights the importance of considering local contexts in science learning and assessment [4].

However, the integration of cultural contexts should not be interpreted as evidence that such contexts automatically improve scientific literacy. In this study, ethnochemistry contexts functioned as stimuli and problem situations. Their effectiveness in improving scientific literacy has not yet been empirically tested.

3.5 Readability and Student Responses in the Limited Trial

The limited trial involved 15 Grade 11 students representing high, moderate, and low academic ability. Data were collected through observations and open-ended interviews conducted after the students completed the instrument.

The observation and interview results indicated that most students were able to understand the instructions, narratives, and ethnochemistry contexts. Students also reported that the Papuan cultural contexts made the questions feel more closely connected to their everyday lives and different from the routine questions they usually encountered.

These findings support the use of local contexts as a means of increasing the meaningfulness of problem situations [7], [26], [28]. The findings are also relevant to scientific literacy assessment, which emphasizes the application of knowledge in context [3], [8], [15].

Nevertheless, contextual familiarity did not automatically make the entire problem-solving process easy. Some students still experienced difficulties in transforming narrative information into calculation procedures, determining relevant data, converting mass to moles, applying coefficient ratios, and interpreting calculation results.

These findings are consistent with previous research on students' difficulties and misconceptions in stoichiometry [21]–[24]. A practical implication is that assessment stimuli should remain contextual, while quantitative information should be presented sufficiently clearly to enable students to move from narrative representations to chemical representations and quantitative calculations.

3.7 Student Responses to AI-Based Feedback

Students responded positively to the AI-based system because they received scores and feedback after completing the questions. The feedback helped students identify which parts of their responses were correct and which aspects required improvement.

This finding is consistent with studies on the use of generative AI to provide feedback on text-based assignments [32]–[34]. Meyer et al. [35] also reported the potential of large language model-generated feedback to support students' revision processes and learning experiences.

However, students' positive responses should not be interpreted as evidence of the accuracy of AI-generated scores. Research on machine-learning scoring indicates that system performance can be influenced by construct characteristics, response types, and scoring criteria [36], [38], [39].

Therefore, AI in this study was positioned as a tool for preliminary scoring and feedback rather than as a replacement for teachers. This position is consistent with the need for human oversight, transparency, and attention to potential algorithmic bias [9], [10], [37], [40].

3.8 Scoring Consistency and Limitations of AI

This study did not examine inter-rater agreement between AI-generated scores and scores assigned by teachers or experts. Therefore, no conclusion can yet be drawn regarding the level of agreement between AI scoring and human raters.

This limitation is important because the content validity of an instrument is distinct from the reliability of its scoring system. An instrument may demonstrate high content validity while its scoring process still requires separate empirical evaluation. Research on the application of machine learning in science assessment emphasizes the importance of comparing automated scoring results with human raters' performance [38], [39].

In future research, students' responses should be independently scored by at least two human raters using the same rubric and subsequently compared with AI-generated scores. For continuous scores, the Intraclass Correlation Coefficient (ICC) could be used, whereas ordinal scores could be analyzed using weighted kappa. The consistency of AI outputs should also be examined through repeated scoring of the same responses under controlled prompting conditions.

In addition, potential algorithmic bias requires attention because algorithm-based decisions may result in differential treatment if the system is not evaluated across diverse groups of students [40]. An ethical framework is also necessary to ensure that AI functions as a support for pedagogical decision-making rather than replacing teachers' professional judgment [37].

3.9 Implications for Instrument Development

Overall, the findings indicate that the AI-supported scientific literacy assessment incorporating Papuan ethnochemistry demonstrated preliminary evidence of high content validity and good readability. The product integrates three main components: scientific literacy, Papuan ethnochemistry contexts, and AI technology.

From an assessment perspective, the open-ended questions provide opportunities for students to explain phenomena, use data, and connect their responses to contextual situations [3], [8], [15]. From a contextual perspective, ethnochemistry provides stimuli that are closely connected to students' experiences [6], [7], [26], [28]–[31]. From a technological perspective, AI has the potential to facilitate more efficient feedback provision [32]–[35].

However, these three components require further empirical investigation. High content validity does not demonstrate the effectiveness of the instrument in improving students' scientific literacy. Similarly, positive student responses to AI do not establish the accuracy of automated scoring. Therefore, the findings should be interpreted as preliminary evidence of the feasibility of the developed product and as a basis for subsequent research involving broader empirical testing and implementation.

Acknowledgments

The authors would like to express their gratitude to the school administrators, chemistry teachers, validators, and Grade 11 students who supported and participated in this study.

References

[1] OECD. (2023). PISA 2022 Results (Volume I): The State of Learning and Equity in Education. Paris: OECD Publishing.

[2] Fuadi, H., Robbia, A. Z., Jamaluddin, J., & Jufri, A. W. (2020). Analisis faktor penyebab rendahnya kemampuan literasi sains peserta didik. Jurnal Ilmiah Profesi Pendidikan, 5(2), 108-116. https://doi.org/10.29303/jipp.v5i2.122

[3] Shwartz, Y., Ben-Zvi, R., & Hofstein, A. (2006). The use of scientific literacy taxonomy for assessing the development of chemical literacy among high-school students. Chemistry Education Research and Practice, 7(4), 203-225.

[4] Listiani, H., Satar, S., Nurbaya, N., & Akobiarek, M. (2024). Profil kemampuan literasi sains (scientific literacy) siswa kelas X SMA YPK Diaspora Kotaraja, Papua. Pendas: Jurnal Ilmiah Pendidikan Dasar, 9(4), 767-778. https://doi.org/10.23969/jp.v9i4.18952

[5] Putri, V. L., & Dj, L. (2022). Description of students learning difficulties in the basic law of chemistry. Jurnal Pijar MIPA, 17(5), 597-603. https://doi.org/10.29303/jpm.v17i5.3855

[6] Azizah, N., & Premono, S. (2021). Identifikasi potensi budaya lokal berbasis etnokimia di Kabupaten Bantul. Journal of Tropical Chemistry Research and Education, 3(1), 53-60. https://doi.org/10.14421/jtcre.2021.31-06

[7] Wardani, L. K., Mulyani, B., Ariani, S. R. D., Yamtinah, S., Masykuri, M., Ulfa, M., & Shidiq, A. S. (2023). Effect of ethnochemistry-based teaching approach on students' learning outcomes. Jurnal Penelitian Pendidikan IPA, 9(12),

11029-11037. https://doi.org/10.29303/jppipa.v9i12.5532

[8] Muniroh, N., Rusilowati, A., & Isnaeni, W. (2022). Instrument development of science literacy assessment with socio-sciences contents in natural science learning for elementary school. Journal of Educational Research and Evaluation, 11(1), 15-22. https://doi.org/10.15294/jere.v11i1.55421

[9] Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., et al. (2021). Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. International Journal of Information Management, 57, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

[10] Khlaif, Z., Odeh, A., & Bsharat, T. R. K. (2024). Generative AI-powered adaptive assessment. In Generative Artificial Intelligence in Education (pp. 157-176). https://doi.org/10.4018/979-8-3693-6397-3.ch007

[11] Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. Educational and Psychological Measurement, 40(4), 955-959. https://doi.org/10.1177/001316448004000419

[12] Musyarofah, A. A. S., Anggraini, A. L., Sudarti, S., Handayani, R. D., Jamhari, M., & Haeruddin, H. (2023). Analysis of the comparison of science literacy skills of students at MTs Nurul Huda Situbondo. EDUKASIA: Jurnal Pendidikan dan Pembelajaran, 4(2), 2507-2516. https://doi.org/10.62775/edukasia.v4i2.619

[13] Morales-Doyle, D., Childress Price, T., & Chappell, M. J. (2019). Chemicals are contaminants too: Teaching appreciation and critique of science in the era of NGSS. Science Education, 103(6), 1347-1366. https://doi.org/10.1002/sce.21546

[14] Branch, R. M. (2009). Instructional Design: The ADDIE Approach. New York: Springer.

[15] Gormally, C., Brickman, P., & Lutz, M. (2012). Developing a Test of Scientific Literacy Skills (TOSLS): Measuring undergraduates' evaluation of scientific information and arguments. CBE-Life Sciences Education, 11(4), 364-377. https://doi.org/10.1187/cbe.12-03-0026

[16] Pratiwi, Z. E., Hidayati, S. N., & Aulia, E. V. (2024). Analisis profil literasi sains siswa kelas VIII di Sidoarjo berdasarkan indikator PISA 2015. Nusantara: Jurnal Pendidikan Indonesia, 4(4), 1131–1140. https://doi.org/10.14421/njpi.2024.v4i4-17

[17] Yusmar, F., & Fadilah, R. E. (2023). Analisis rendahnya literasi sains peserta didik Indonesia: Hasil PISA dan faktor penyebab. LENSA (Lentera Sains): Jurnal Pendidikan IPA, 13(1), 11–19. https://doi.org/10.24929/lensa.v13i1.283

[18] Sari, D. N. A., Rusilowati, A., & Nuswowati, M. (2017). Pengaruh pembelajaran berbasis proyek terhadap kemampuan literasi sains siswa. PSEJ (Pancasakti Science Education Journal), 2(2), 114–120. https://doi.org/10.24905/psej.v2i2.741

[19] Dinata, A. N., Adisendjaja, Y. H., & Amprasto, A. (2018). Pengaruh field trip terhadap kemampuan literasi sains dan sikap terhadap sains siswa SMA pada materi ekosistem. Assimilation: Indonesian Journal of Biology Education, 1(1), 8–13. https://doi.org/10.17509/aijbe.v1i1.11449

[20] OECD. (2023). PISA 2022 Assessment and Analytical Framework. Paris: OECD Publishing. https://doi.org/10.1787/dfe0bf9c-en

[21] Lestari, E. A., Harjito, Susilaningsih, E., & Wijayati, N. (2021). Analisis miskonsepsi menggunakan tes diagnosa three-tier multiple choice pada materi stoikiometri. Jurnal Inovasi Pendidikan Kimia, 15(2), 2824–2830. https://doi.org/10.15294/jipk.v15i2.29171

[22] Sidauruk, S. (2005). Miskonsepsi stoikiometri pada siswa SMA. Jurnal Penelitian dan Evaluasi Pendidikan, 7(2). https://doi.org/10.21831/pep.v7i2.2024

[23] Sari, S. F. I., Fatah, A. H., & Anggraeni, M. E. (2020). Analisis materi ajar kimia SMA/MA kelas X pada konsep stoikiometri. Jurnal Ilmiah Kanderang Tingang, 11(1), 139–151. https://doi.org/10.37304/jikt.v11i1.82

[24] Basuki, K. H. (2020). Analisis kesalahan mahasiswa dalam menyelesaikan soal stoikiometri. Jurnal Ilmiah Wahana Pendidikan, 6(4), 641–651. https://doi.org/10.5281/zenodo.4298027

[25] Sariati, N. K., Suardana, I. N., & Wiratini, N. M. (2020). Analisis kesulitan belajar kimia siswa kelas XI pada materi larutan penyangga. Jurnal Ilmiah Pendidikan dan Pembelajaran, 4(1), 100–108. https://doi.org/10.23887/jipp.v4i1.15469

[26] Jofrishal, & Seprianto. (2020). Implementasi modul kimia pangan melalui pendekatan etnokimia di SMK Negeri Aceh Timur program keahlian tata boga. Jurnal IPA &

Pembelajaran IPA, 4(2), 168–177. https://doi.org/10.24815/jipi.v4i2.17262

[27] Jumriani, Mutiani, Putra, M. A. H., Syaharuddin, & Abbas, E. W. (2021). The urgency of local wisdom content in social studies learning: Literature review. The Innovation of Social Studies Journal, 2(2), 103–109. https://doi.org/10.20527/iis.v2i2.3076

[28] Wahyudiati, D., & Fitriani. (2021). Etnokimia: Eksplorasi potensi kearifan lokal Sasak sebagai sumber belajar kimia. Jurnal Pendidikan Kimia Indonesia, 5(2), 102–111. https://doi.org/10.23887/jpk.v5i2.38537

[29] Aldiansyah, Pasa, J. I., Muttaqin, M. R., Awaliyah, N. N., & Erika, F. (2023). Literatur review: Keterkaitan pembelajaran kimia terhadap pendekatan etnokimia di Indonesia. CHEDS: Journal of Chemistry, Education, and Science, 7(2), 238–246. https://doi.org/10.30743/cheds.v7i2.8416

[30] Zahro, A. N., Sumarni, W., & Kurniawan, C. (2023). Tradition of making lontong tuyuhan in Rembang Regency as a science learning resource. Jurnal Penelitian Pendidikan IPA, 9(1), 207–215. https://doi.org/10.29303/jppipa.v9i1.2197

[31] Gazali, Z., & Andriani, R. (2025). Integrasi etnokimia dalam pengajaran IPA: Analisis literatur sistematis pada pendekatan berbasis kearifan lokal di SMA. Jurnal Pendidikan MIPA, 15(4), 1843–1853. https://doi.org/10.37630/jpm.v15i4.3783

[32] Lee, S. S., & Moore, R. L. (2024). Harnessing generative AI (GenAI) for automated feedback in higher education: A systematic review. Online Learning, 28(3), 82–104. https://doi.org/10.24059/olj.v28i3.4593

[33] Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? International Journal of Educational Technology in Higher Education, 21(1), 23. https://doi.org/10.1186/s41239-024-00455-4

[34] Naseer, F., Khalid, M. U., Ayub, N., Rasool, A., Abbas, T., & Afzal, M. W. (2024). Automated assessment and feedback in higher education using generative AI. In R. C. Sharma & A. Bozkurt (Eds.), Transforming Education With Generative AI: Prompt Engineering and Synthetic Content Creation (pp. 433–461). Hershey, PA: IGI Global. https://doi.org/10.4018/979-8-3693-1351-0.ch021

[35] Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. Computers and Education: Artificial Intelligence, 6, 100199. https://doi.org/10.1016/j.caeai.2024.100199

[36] Haudek, K. C., & Zhai, X. (2024). Examining the effect of assessment construct characteristics on machine learning scoring of scientific argumentation. International Journal of Artificial Intelligence in Education, 34(4), 1482–1509. https://doi.org/10.1007/s40593-023-00385-8

[37] Swindell, A., Greeley, L., Farag, A., & Verdone, B. (2024). Against artificial education: Towards an ethical framework for generative artificial intelligence (GenAI) use in education. Online Learning, 28(2), 7–28. https://doi.org/10.24059/olj.v28i2.4438

[38] Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: A systematic review. Studies in Science Education, 56(1), 111–151. https://doi.org/10.1080/03057267.2020.1735757

[39] Beggrow, E. P., Ha, M., Nehm, R. H., Pearl, D., & Boone, W. J. (2014). Assessing scientific practices using machine-learning methods: How closely do they match clinical interview performance? Journal of Science Education and Technology, 23(1), 160–182. https://doi.org/10.1007/s10956-013-9461-9

[40] Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. International Journal of Artificial Intelligence in Education, 32(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9