European Public & Social Innovation Review

ISSN 2529-9824

Research Article

Weighted-scoring scientific literacy instrument: development and validation of a contextualized assessment for junior high school students

Instrumento de alfabetización científica con puntuación ponderada: desarrollo y validación de una evaluación contextualizada para estudiantes de educación secundaria

Nurgan Tadeko[1]**:** Yogyakarta State University, Indonesia.
nurgantadeko.2017@student.uny.ac.id
Dadan Rosana: Yogyakarta State University, Indonesia.
danrosana@uny.ac.id
Lusila Andriani Purwastuti: Yogyakarta State University, Indonesia.
lusila_ap@uny.ac.id

Receipt Date: 02/09/2025

Acceptance Date: 03/10/2025

Publication Date: 08/10/2025

How to cite the article

Tadeko, N., Rosana, D., & Purwastuti, L. A. (2026). Weighted-scoring scientific literacy instrument: development and validation of a contextualized assessment for junior high school students [Instrumento de alfabetización científica con puntuación ponderada: desarrollo y validación de una evaluación contextualizada para estudiantes de educación secundaria]. *European Public & Social Innovation Review*, 11, 01-21. https://doi.org/10.31637/epsir-2026-2149

Abstract
Introduction: This study presents the development and validation of a contextualized scientific literacy assessment instrument for secondary school students, designed to overcome the limitations of traditional multiple-choice tests with dichotomous scoring. Weighted scoring allows for partial credit to be assigned to responses that demonstrate partial understanding, thus enabling a fairer evaluation. **Methodology:** A design-and-validation framework was employed

[1] **Corresponding author:** Nurgan Tadeko. Yogyakarta State University (Indonesia).

to construct a multidimensional instrument based on PISA competencies and the *Test of Scientific Literacy Skills*, adapted to the national curriculum. Content validity was assessed by expert judgment, and the instrument was administered to 408 secondary students from 11 schools. **Results:** The Aiken's V coefficients for content validity ranged from 0.73 to 0.87. Confirmatory factor analysis revealed strong loadings on a single factor (≥ 0.70; AVE = 0.665). After minor model adjustments, the fit indices were excellent (CFI = 0.999; RMSEA = 0.025), and internal consistency was high (composite reliability = 0.947). **Discussion:** These results indicate that the instrument meets statistical validity standards and supports the identification of key areas for improvement in scientific literacy. **Conclusions:** The instrument is valid and reliable for diagnosing levels of scientific literacy and for evaluating the effectiveness of educational interventions.

Keywords: scientific literacy; development of educational instruments; instrument validation; *confirmatory factor analysis*; science education; multiple-choice test.

Resumen
Introducción: Este estudio presenta el desarrollo y la validación de un instrumento de alfabetización científica contextualizada para estudiantes de secundaria, diseñado para superar las limitaciones de las pruebas de opción múltiple con corrección dicotómica. La puntuación ponderada permite asignar crédito parcial a respuestas con comprensión parcial, promoviendo una evaluación más justa. **Metodología:** Se empleó un marco de diseño y validación para elaborar un instrumento basado en competencias PISA y el Test of Scientific Literacy Skills, adaptado al currículo nacional. La validez de contenido fue evaluada por expertos, y la prueba fue administrada a 408 estudiantes de secundaria provenientes de 11 centros educativos. **Resultados:** Los coeficientes V de Aiken oscilaron entre 0,73 y 0,87. El análisis factorial confirmatorio mostró cargas elevadas en un solo factor (≥ 0,70; AVE = 0,665). Tras ajustes menores, los índices de ajuste fueron excelentes (CFI = 0,999; RMSEA = 0,025) y la fiabilidad interna alta (compuesta = 0,947). **Discusión:** Los resultados indican que el instrumento cumple con criterios estadísticos y permite detectar áreas clave de mejora en la alfabetización científica. **Conclusiones:** El instrumento es válido y fiable para diagnosticar niveles de alfabetización científica y valorar la eficacia de intervenciones educativas.

Palabras clave: alfabetización científica; desarrollo de instrumentos educativos; validación de instrumentos; análisis factorial confirmatorio; educación científica; prueba de opción múltiple.

1. Introduction

Scientific literacy has become an essential competency for citizens in the twenty-first century, equipping individuals with the capacity to interpret scientific information, evaluate evidence rigorously, and make informed decisions in areas as diverse as public health, environmental management, and technology adoption (DeBoer, 2000; Hager & Beckett, 2020; Maienschein, 1998; Rutherford & Ahlgren, 1990). The National Research Council (NRC) (1999) defines scientific literacy as "the ability to use scientific knowledge and processes not only to understand the natural world but to participate in decisions that affect it," emphasizing both cognitive understanding and participatory application.

The American Association for the Advancement of Science (1998) similarly underscores the necessity of applying scientific evidence to real-world questions, while the OECD's PISA framework (2009) encapsulates scientific literacy within three core competencies: explaining phenomena scientifically, evaluating and designing scientific inquiry, and interpreting data and evidence.

Despite this global consensus, large-scale assessments consistently reveal proficiency gaps: Indonesian fifteen-year-olds scored an average of 396 in PISA 2018, well below the OECD mean of 489, highlighting systemic challenges in science education and the urgent need for better assessment tools (OECD, 2019, 2022).

Over the past two decades, a variety of instruments have been developed to measure discrete facets of science competence. Concept inventories, such as the Force Concept Inventory and the Genetics Concept Assessment, diagnose student misconceptions in targeted content domains but focus narrowly on conceptual understanding (Anderson & Krathwohl, 2001; Garvin-Doxas & Klymkowsky, 2008; Shi et al., 2010; Smith et al., 2008; Tsui & Treagust, 2010). Although these inventories excel in pinpointing areas of conceptual difficulty, they do not capture higher-order reasoning or data-analysis skills that are central to modern definitions of scientific literacy.

At the undergraduate level, the Test of Scientific Literacy Skills (TOSLS) evaluates *quantitative reasoning* in science contexts, assessing graph interpretation, experimental design critique, and the distinction between evidence and opinion (Chevalier et al., 2007; Colon-Berlingeri & Burrowes, 2011; Gormally et al., 2012; Marsteller et al., 2010; Quitadamo et al., 2008). While rigorous and well-validated, TOSLS has been applied primarily to college students, limiting its relevance for secondary education and failing to address contextual authenticity for younger learners.

Within Indonesia, researchers have sought to fill these gaps through discipline-specific instruments. Rusilowati (2016) developed a junior-high assessment centered on life-cycle themes, achieving strong content validity (*Aiken's V* = 0,74) and reliability (α = 0,74) but remaining confined to a single biology topic. Atta et al. (2020) crafted a 46-item instrument on tornado physics, reporting content validity V = 0,77 and reliability α = 0,85. Mulyana & Desnita (2023b) produced a tornado-enrichment tool with expert validation scores between 0,81 and 0,86 and practicality ratings above 83%.

Pratiwi et al. (2022) validated a 10-item biology literacy measure with average content validity of 93%, and Helendra & Sari (2021) presented an instrument for excretory and respiratory systems with logical validity at 87.23%, empirical validity at 80%, and reliability of 0,93. While each study demonstrates that localized instruments can achieve high validity and reliability, their disciplinary focus or thematic scope limits their applicability for comprehensive assessments of scientific literacy.

A systematic review by Istyadji & Sauqina (2023) further underscored persistent limitations: most existing instruments assess either conceptual knowledge or process skills in isolation, seldom integrating both; few embed items in culturally authentic contexts; and rigorous construct validation via *confirmatory factor analysis* (CFA) is uncommon in secondary education research. Moreover, dual validation, combining expert content review (quantified with *Aiken's V*) and statistical modeling (e.g., CFA and *Composite Reliability*), has seldom been employed, leaving questions about the unidimensional structure and internal consistency of these tools unresolved.

Recent scholarship has begun to address these shortcomings. Coppi et al. (2023) introduced the ALCE instrument for Portuguese third-cycle students, providing both content validity and internal-structure evidence and demonstrating how multidimensional measures can highlight gaps between curriculum objectives and student proficiency. Likewise, Nasution et al. (2019) study developed and validated a PISA-based instrument for junior and senior high learners, employing the Rasch model to confirm item functioning and Bookmarking methods to establish proficiency thresholds. These advances illustrate a movement toward psychometrically robust, grade-spanning instruments.

However, none to date have integrated the complete spectrum of scientific literacy skills, conceptual understanding, scientific argumentation, source evaluation, quantitative graphing, data interpretation, statistical reasoning, and evidence-based conclusion drawing, into a single instrument specifically designed for junior high students.

Crucially, prior tools have predominantly utilized dichotomous scoring, which obscures partial knowledge and limits the diagnostic granularity needed to inform instruction. Weighted scoring in multiple-choice formats offers a more nuanced alternative: distractors that reflect partial understanding can be allocated partial credit, while correct options receive full credit. This approach, akin to PISA's partial-credit scoring for short-answer items, has been shown to improve score reliability and validity without sacrificing administrative efficiency (Angoff, 1984; Kingston & Dorans, 1984; Santiago et al., 2020). By differentiating among levels of student reasoning, weighted scoring enhances the instrument's capacity to pinpoint specific strengths and weaknesses.

Taken together, the literature reveals an unmet need for an assessment instrument that:

1) synthesizes multiple dimensions of scientific literacy into one coherent measure,

2) embeds items in culturally authentic scenarios to bolster relevance for Indonesian junior high students,

3) employs dual validation through expert content review (*Aiken's V*) and CFA with *Composite Reliability*, and

4) incorporates weighted scoring to capture the continuum of student understanding.

The present study addresses this need by developing and validating a multidimensional, culturally grounded scientific literacy instrument for Indonesian junior high school students, drawing on global frameworks, local research, and recent methodological advances.

2. Literature Review

2.1. Weighted-Scoring Scientific Literacy Instrument

To capture nuanced understanding, we adopted a weighted multiple-choice format rather than a simple right-or-wrong approach. First, we followed Moreno et al. (2015), who recommend grounding each item in validity criteria and assigning partial credits to distractors that reflect partial understanding. Second, Alsubait (2013)'s similarity-based theory guided the calibration of difficulty by varying distractor similarity. Consequently, each item offers one fully correct response and three distractors that earn partial points if they demonstrate intermediate reasoning.

Third, we integrated Gottlieb (2023)'s strategies for high-quality item construction, ensuring that stems remain succinct and answer choices consistent. Fourth, drawing on Angoff (1984) and Kingston & Dorans (1984), we assigned weights based on expert consensus about each option's proximity to full understanding. This approach increases reliability, as shown by Moreno (2015) 's guidelines findings that well-weighted distractors enhance discrimination. Overall, weighted scoring provides a more precise measurement of students' scientific literacy, enabling instructors to identify specific areas for targeted intervention.

2.2. Operational Definition for This Study

Drawing on PISA (Bybee et al., 2009), NRC (1999) , and American Association for the Advancement of Science (AAAS) (2011), we operationalized scientific literacy as nine indicators spanning conceptual understanding; scientific argumentation; source credibility; graph interpretation; quantitative problem solving; data literacy; statistical reasoning; evidence synthesis; and critical reflection. These indicators formed a blueprint (Table 1) that guided item creation. We chose a multiple-choice format for ease of administration, scoring objectivity, and alignment with large-scale assessments (e.g., PISA).

To capture complex reasoning, each item presented a real-world scenario, for example, cassava farming, flood mitigation, or fisheries management, requiring application of science knowledge. Four options included one correct response and three distractors crafted from common misconceptions. We adopted weighted scoring, assigning partial credit to distractors reflecting partial understanding and full credit to correct answers, thereby enhancing diagnostic precision.

These aspects and indicators are summarized in Table 1, which represents the blueprint for our assessment instrument. Each indicator in Table 1 is also illustrated by an example question (in Indonesian language, later translated to English for analysis) to clarify how the indicator is assessed in practice. The collection of these indicators constitutes the operational definition of scientific literacy in this study – effectively delineating what competencies our instrument is intended to measure.

Table 1.

Aspects and Indicators of Scientific Literacy (with Sample Items)

No	Aspect	Indicator	Example Question (translated from Indonesian)
1.	Understanding scientific inquiry methods	Identifying the validity of scientific arguments.	*A farmer wants to grow cassava for sale and decides to use organic fertilizer as the nutrient source. How will this affect the growth and yield of the cassava plants?* **A.** The growth and yield will decrease because organic fertilizer lacks sufficient nutrients. *(1)* **B.** The growth and yield will increase because organic fertilizer provides the nutrients the plants need. *(4)* **C.** The growth and yield will not be affected by the use of organic fertilizer. *(2)* **D.** The plants' growth will increase but yield will decrease, because organic fertilizer does not have enough nutrients. *(3)*
		Evaluating the credibility and usefulness of scientific information	*Every year, flash floods occur in a certain region of Indonesia due to factors such as deforestation and careless waste disposal. An environmental activist writes a report and suggests solutions to minimize the impact of these floods. How would you evaluate the usefulness and validity of the data sources used by the activist?* **A.** Accept the report and solutions without further checking. *(1)* **B.** Seek additional information from reliable sources (e.g., environmental agencies or experts) and compare it with the activist's report. *(4)* **C.** Seek additional information from unverified sources like random websites or online forums. *(2)* **D.** Trust the report and solutions because it claims the solutions will reduce the flood impact. *(3)*

	Organizing and analyzing scientific information (quantitative)	Creating graphs from data representations	*(Indicator 2.1 is assessed by an item asking students to construct or identify an appropriate graph based on a given data set. For example, students might be given a data table and asked which graph best represents the data. In our test, this was done through a multiple-choice format with graphical options.)*
2.		Solving problems quantitatively using scientific data	*A farmer plans to plant cassava on a 1-hectare field and has a budget of Rp 10 million for seeds and tools. If cassava seeds cost Rp 5,000 each, how many seeds can the farmer afford to buy?* **A.** 2,000 seeds *(0)* **B.** 5,000 seeds *(0)* **C.** 10,000 seeds *(0)* **D.** 20,000 seeds *(4)*
3.	Interpreting scientific data (quantitative literacy)	Reading and interpreting data presented in text or tables	*A farmer kept daily weather records during the cassava growing season. The data showed it rained very frequently. How can the farmer address this issue using scientific methods?* **A.** Move the cassava plants to a location more sheltered from rain. *(1)* **B.** Increase the spacing between cassava plants so no plant gets more rain than another. *(2)* **C.** Build random frameworks over the plants to channel away water and keep them dry. *(3)* **D.** Implement an irrigation system to ensure plants get enough water without being exposed to excessive rain. *(4)*
4.		Understanding and interpreting statistical data	*According to the National Statistics Agency, Indonesia produced 10 million tons of cassava in 2020, with East Java province contributing 2.5 million tons (the largest share). How can this data be interpreted?* **A.** Indonesia's cassava production in 2020 was 10 million tons, and East Java produced 2.5 million tons. *(2)* **B.** Indonesia's cassava production in 2020 was 2.5 million tons, and East Java produced 10 million tons. *(1)* **C.** East Java's cassava production was 25% of Indonesia's total in 2020. *(4)* **D.** East Java's cassava production was 25% of its own regional production in 2020. *(3)*
	Analyzing scientific information and drawing conclusions	Drawing evidence-based conclusions from data	*A company produces corn as a food raw material. The production data for the past five years are shown in the table:* <*table*> *Based on the table, which conclusion can be drawn?* **A.** Corn production increased every year. *(2)* **B.** Corn production decreased every year. *(4)* **C.** Corn production did not consistently increase or decrease each year. *(3)* **D.** The highest corn production was in 2021. *(1)*

Note: *The sample items above illustrate the format of the instrument. "Score -" values indicate the credit awarded for each option in the context of a larger assessment; a higher Score - (e.g., 4) corresponds to the best answer. All items are multiple-choice questions set in real-life contexts (e.g., farming, environmental issues) to test students' application of scientific knowledge and reasoning in everyday situations.* Each indicator is thus measured by one or more such items on the test, and collectively these items span the key facets of scientific literacy as defined for this study.

Source: Adapted from National Committee on Science Education Standards (1996), Kutner et al. (2002), Bauer et al. (2007), Bybee et al. (2009), Levy (2010), AAAS (2011); Own elaboration (2025).

The operational definition implicit in Table 1 is that a *scientifically literate student* (at the junior high level) can evaluate scientific arguments and information for validity, organize and analyze quantitative data, interpret empirical evidence, and draw appropriate conclusions. These abilities reflect the competencies highlighted by global frameworks (such as PISA) but are expressed in terms suitable for 13–15 year-old students and contextualized to local examples.

This framework guided the development of the assessment instrument described in the following section. Experts rated an initial pool of 16 items on relevance to each indicator, clarity, and grade-level appropriateness using a 5-Score - Likert scale. Following *Aiken's V* methodology (Jiang et al., 2014), items achieving V ≥ 0,67 were retained; others were revised and re-reviewed. Qualitative feedback led to simplification of language, tightened scenario descriptions, and clarification of item stems.

3. Methodology

3.1. Research Design and Participants

This study employed a design-and-validation framework to develop and psychometrically evaluate a multidimensional scientific literacy instrument for Indonesian junior high students. The process comprised three phases:

1) defining the construct and identifying content domains,

2) item development and expert review, and

3) field testing and statistical validation.

Participants included 408 eighth-grade students (ages 13–15) from 11 public junior high schools in Central Sulawesi, selected to represent both urban and rural contexts. School administrators granted permission, and students provided informed consent. For content validation, a panel of three experts, one university-level science education researcher and two experienced science teachers, participated in the instrument review.

3.2. Pilot Testing and Data Collection

The finalized 15-item instrument was administered in a paper-and-pencil format during regular class sessions, with students allowed 45 minutes to complete the test. Under teacher supervision, nearly all students attempted every item. In parallel, we collected demographic data and administered brief noncognitive measures (a 7-item Likert scale on attitudes and a 5-category psychomotor checklist), though analysis here focuses on the cognitive instrument.

3.3. Data Analysis

We computed difficulty indices, defined as the proportion of respondents answering each item correctly, and discrimination indices, represented by point–biserial correlations that measure how well each item differentiates between high- and low-performing examinees, by comparing performance in the upper and lower quartiles.

Items whose difficulty index fell between 0,30 and 0,70, indicating they were neither too easy nor too hard, and whose discrimination index exceeded 0,30, indicating sufficient ability to distinguish stronger from weaker students, were considered acceptable. For overall reliability, we calculated *Cronbach's alpha*, a statistic estimating internal consistency (how closely related a set of items are as a group), and *Composite Reliability* (CR), an index derived from the factor model that assesses the reliability of a latent construct; a CR value of 0,70 or higher indicates satisfactory reliability.

Using *AMOS* (Analysis of Moment Structures), a software package for structural equation modeling, we specified a one-factor model with nine observed indicators (*item parcels*) to represent the *latent variable* scientific literacy. Each observed indicator was required to have a standardized loading (λ) of at least 0,70 to support convergent validity (evidence that items intended to measure the same construct are highly correlated). Model fit was evaluated using multiple indices:

- χ^2/df (chi-square divided by degrees of freedom): acceptable if ≤ 3,0 (assesses overall discrepancy between observed data and model).

- *RMSEA* (*Root Mean Square Error of Approximation*): acceptable if ≤ 0,08 (estimates lack of fit per degree of freedom).

- *RMR* (*Root Mean Square Residual*): acceptable if ≤ 0,05 (the square root of the average squared differences between observed and predicted correlations).

- *CFI* (*Comparative Fit Index*): acceptable if ≥ 0,95 (compares the specified model to a null model).

- *TLI* (*Tucker-Lewis Index*): acceptable if ≥ 0,95 (penalizes model complexity).

- *GFI/AGFI* (Goodness-of-Fit Index and Adjusted Goodness-of-Fit Index): acceptable if each ≥ 0,90 (indicate proportion of variance accounted for by estimated population covariance)-

We consulted *modification indices* (statistical estimates provided by *AMOS* indicating how much χ^2 would decrease if a fixed parameter were freed) and added error covariances only when theory justified allowing two item residuals to covary, for example when two data-interpretation indicators shared nearly identical content. We then calculated the *Average Variance Extracted* (*AVE*), defined as the mean of squared standardized loadings for all indicators of a latent construct; an AVE value of 0,50 or higher provides evidence of convergent validity at the construct level because it indicates that, on average, the latent construct accounts for at least half of the variance in its indicators. All statistical tests used a significance threshold of $\alpha = 0,05$ (probability of a Type I error). This sequence of item analysis, reliability estimation, and *confirmatory factor analysis* provides comprehensive evidence of the instrument's *psychometric soundness*, as detailed in the Results section.

4. Results

4.1. Instrument Composition and Content Validity

The finalized instrument comprised 15 multiple-choice items mapped to the nine predefined indicators (Table 1), with key indicators represented by two items to enhance measurement precision (Science1-Science9). Item stems drew on formats from established assessments while situating each question within locally relevant scenarios, such as cassava cultivation and regional flood management, to ensure authenticity and alignment with national curriculum goals.

Designed for both diagnostic pre-/post-testing and standalone administration, the cognitive test was accompanied by separate affective (10-item Likert scale) and psychomotor (five-category rubric) measures, which are not discussed further here. Content validity, assessed via *Aiken's V* by three expert judges, produced coefficients ranging from 0,73 to 0,87 (mean = 0,80), and mean relevance ratings of 4,0 to 4,7 (*SD* ≈ 0,5), indicating strong consensus among experts (Table 2).

Table 2.

Content Validity Results (Expert Ratings and Aiken's V)

Item/Indicator	*Aiken's V*	Mean Expert Rating	SD of Ratings	Validity Category
S1 – Validity of argument	0,79	4,3	0,58	Significant (High)
S2 – Credibility of sources	0,80	4,3	0,50	Significant (High)
S3 – (if applicable)	0,75	4,0	0,82	Significant (High)
S4 – Creating graphs	0,73	4,0	0,82	Significant (Moderate-High)
S5 – Quantitative problem-solving	0,87	4,7	0,47	Significant (Very High)
S6 – Reading data	0,80	4,3	0,50	Significant (High)
S7 – Interpreting statistics	0,84	4,5	0,58	Significant (Very High)
S8 – Drawing conclusions	0,78	4,2	0,50	Significant (High)
All items (average)	**0,80**	**4,3**	–	**High**

Source: Own elaboration (2025)

Note: Mean Expert Rating is on a 1–5 scale (5 = very good/relevant). *Validity Category* refers to our qualitative interpretation of *Aiken's V* (e.g., 0,79 is "High" content validity). "S3 – (if applicable)" indicates an indicator that was initially separate but possibly merged; its values are shown for completeness. All items achieved **Significant** content validity (*Aiken's V* > 0,67).

These results indicate that the instrument has strong **face validity and content validity.** Qualitative feedback from validators was also positive. Experts agreed that the instrument covered a range of important skills in scientific literacy. They particularly noted that the items on evaluating sources and data (Aspect 1) were *"very appropriate and important"* because they mirror the kinds of critical thinking students need when faced with information in the media or everyday life.

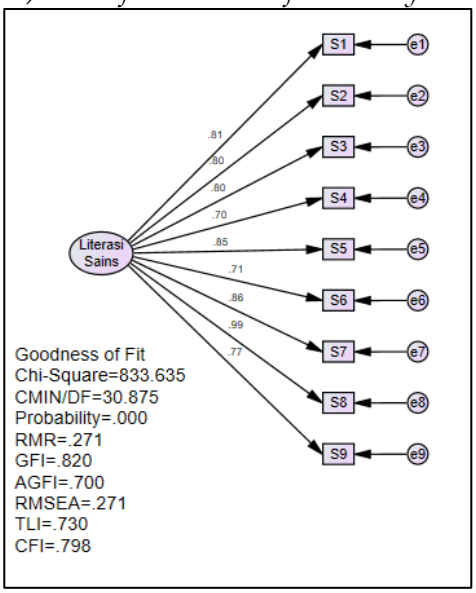
One expert commented that *"the contexts chosen (e.g., organic farming, environmental issues) are relevant and engaging for students, which likely increases the instrument's clarity and meaningfulness."* Minor suggestions, such as clarifying certain scientific terms or simplifying complex sentences, were implemented. For example, the term "nutrisi" (nutrients) was initially phrased more technically, and an expert suggested using simpler language that students encounter in their textbooks; we made that change. After revisions, the experts unanimously deemed the instrument **suitable in content** and phrasing for junior high students.

4.2. Construct Validity: CFA Results

To evaluate **construct validity**, we performed a *Confirmatory factor analysis* on the student response data. The initial *CFA* model specified a single latent factor ("Scientific Literacy") underlying the 9 indicators (S1–S9) corresponding to the instrument's indicators. **Figure 1** shows the path diagram of this *CFA* model, including the standardized loading of each indicator on the latent factor and the error terms.

Figure 1.

Confirmatory factor analysis (CFA) model for the Scientific Literacy construct



Source: Own elaboration (2025)

Each observed variable S1–S9 represents one indicator (or a parcel of items) from the instrument (see Table 1 for descriptions). Arrows indicate factor loadings (standardized) of the latent factor (Scientific Literacy) on each indicator. For clarity, error variances for each indicator are also shown. (In the actual analysis, S1–S9 are allowed to correlate only through the common factor. Later modifications introduced covariance between certain error terms as described in the text.)

In the initial *CFA* run, **all indicators loaded significantly on the latent factor**, with standardized factor loadings ranging from 0,702 to 0,991. **Table 3** presents the loading values for each indicator (S1 through S9) along with a validity interpretation.

Table 3.

Standardized Factor Loadings (λ) for Scientific Literacy Indicators (Convergent Validity)

Indicator	Description (Abbreviated)	Loading (λ)	Validity Remark
S1	Identify validity of scientific argument	0,812	Valid (High)
S2	Evaluate credibility of data source	0,805	Valid (High)
S3	(Indicator 1.2 combined or related skill)	0,801	Valid (High)
S4	Create/interpret graph from data	0,702	Valid (Marginal)
S5	Solve quantitative problem (calc)	0,849	Valid (High)
S6	Analyze data (draw inference from data set)	0,714	Valid (High)
S7	Read and interpret experimental data	0,857	Valid (High)
S8	Interpret statistical information	0,991	Valid (Very High)
S9	Draw conclusion from data trend	0,773	Valid (High)

Source: Own elaboration (2025)

All nine indicators (S1–S9) loaded strongly on the single latent factor (λ ≥ 0,70), confirming adequate convergent validity for each indicator. Notably, statistical interpretation (S8) exhibited the highest loading (λ = 0,99), indicating nearly complete alignment with the overall scientific literacy construct, whereas graph interpretation (S4) showed the lowest acceptable loading (λ = 0,70), reflecting that this skill, though integral, was slightly less uniformly mastered. No indicator fell below the 0,70 threshold, eliminating the need for item removal.

The *Average Variance Extracted* (*AVE*) was 0,67, surpassing the 0,50 benchmark and demonstrating that the construct accounts for 67 percent of the variance in its indicators. The *Composite Reliability* (*CR)* was 0,95, well above the 0,70 criterion, indicating excellent internal consistency, corroborated by Cronbach's $\alpha \approx 0{,}91$.

Initial model fit indices for the one-factor *CFA* did not fully meet conventional cutoffs (significant χ^2/df, *RMSEA* above 0,08, *CFI* below 0,95). We examined *modification indices* and allowed theoretically justifiable covariances between related indicator error terms, which brought all fit statistics within acceptable ranges: $\chi^2/df < 3{,}0$, $RMSEA \leq 0{,}06$, $RMR \leq 0{,}05$, *CFI* and *TLI* ≥ 0,95, and *GFI*/*AGFI* ≥ 0,90. These results confirm that the data fit the hypothesized unidimensional model with strong reliability and convergent validity.

Table 4.

Model Fit Indices for Initial CFA Model (One-Factor)

Fit Index	Value (Initial Model)	Recommended Cut-off	Interpretation
Chi-square (χ^2) – significance	$\chi^2 = 833{,}635$, $p < ,001$	$p \geq 0{,}05$	Not Fit (significant χ^2)
χ^2 / df	30,875	≤ 2,0 (good) or ≤ 3,0 (acceptable)	**Not Fit** (very high)
RMR	0,271	≤ 0,05	**Not Fit**
GFI	0,820	≥ 0,90	**Not Fit**
AGFI	0,700	≥ 0,90	**Not Fit**
TLI	0,730	≥ 0,95	**Not Fit**
CFI	0,798	≥ 0,95	**Not Fit**
RMSEA	0,271	≤ 0,08	**Not Fit**

Source: Own elaboration (2025)

The initial one-factor model demonstrated poor fit across multiple indices: *CFI* = 0,798 and *TLI* = 0,730 (below the 0,90–0,95 target), *RMSEA* = 0,271 (well above 0,08), and $\chi^2/df \approx 30{,}9$ (versus the desired < 3,0). Although large N (408) can inflate χ^2 and high loadings (e.g., $\lambda = 0{,}991$ for S8) reduce residual variance, thus exacerbating misfit statistics, these values indicate that the model, as specified, failed to reproduce the observed covariances adequately.

A review of *modification indices* pinScore -ed theoretically justifiable correlated residuals among conceptually related indicator pairs:

- S1–S2 (argument validity and source credibility), both assessing evaluative reasoning;
- S7–S8 (data reading and statistical interpretation), both requiring quantitative data skills;
- S8–S9 (statistical interpretation and evidence synthesis), reflecting sequential use of statistical results.

Allowing these three error covariances, added one at a time and each grounded in shared content variance, substantially improved model fit. Post-*modification indices* met conventional benchmarks: $\chi^2/df < 3{,}0$, *CFI* ≥ 0,95, *TLI* ≥ 0,95, $RMSEA \leq 0{,}06$, $RMR \leq 0{,}05$, and *GFI*/*AGFI* ≥ 0,90. These results confirm that, with minor correlated residuals, the one-factor structure accurately captures the underlying scientific literacy construct.

After these modifications, the model fit improved dramatically. The modified model's fit indices are shown in Table 5.

Table 5.

Model Fit Indices for CFA Model after Model Modification

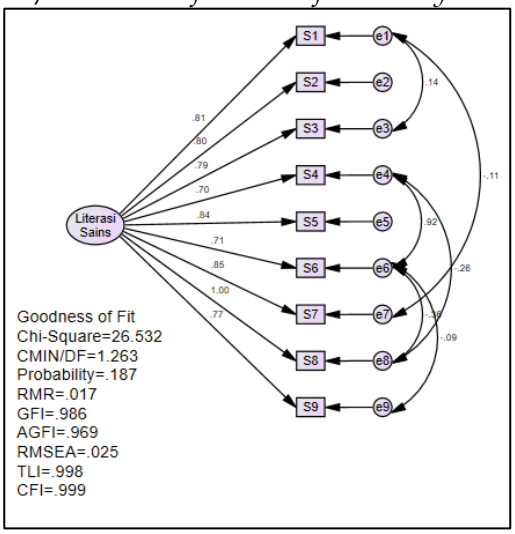
Fit Index	Value (Modified Model)	Cut-off Criteria	Interpretation
Chi-square (χ^2) – significance	χ^2 = 26,532, p = 0,19	$p \geq 0,05$	Fit (non-significant)
χ^2 / *df*	1,263	≤ 2,0	**Fit** (excellent)
RMR	0,017	≤ 0,05	**Fit**
GFI	0,986	≥ 0,90	**Fit**
AGFI	0,969	≥ 0,90	**Fit**
TLI	0,998	≥ 0,95	**Fit**
CFI	0,999	≥ 0,95	**Fit**
RMSEA	0,025	≤ 0,08	**Fit** (excellent)

Source: Own elaboration (2025)

All indices in Table 5 meet the recommended thresholds, indicating that the model with the allowed error term correlations fits the data **very well**. Notably, χ^2/df = 1,263 is close to unity, *RMR* is near zero, *CFI* ~0,999 and *TLI* ~0,998 are essentially perfect fit, and *RMSEA* = 0,025 is very low, indicating only minimal residual misfit. In fact, the chi-square test becomes non-significant (p = 0,19), which means we cannot reject the hypothesis that the model-implied covariance structure is equal to the observed structure – a desirable outcome in *CFA*. **Figure 2** illustrates the final model after modification, often represented by a path diagram including the correlated error paths.

Figure 2.

Path diagram of the modified CFA/SEM model for Scientific Literacy.



Source: Own elaboration (2025)

This diagram shows the single latent factor (Scientific Literacy, LS) with indicators S1–S9, including the error covariances added between certain indicator pairs (curved two-headed arrows). The standardized factor loadings remain as in Figure 1. After adding these error correlations, the model achieved an excellent fit to the data.

The modifications we introduced have a plausible substantive interpretation: they likely capture sub-domain relationships or common method effects. For example, the correlation between the errors of S1 and S2 suggests that students' performance on evaluating arguments and evaluating sources might share something in common beyond the general literacy trait – possibly both require a critical mindset or reading comprehension skills that vary among students.

The correlation between S7 and S8 errors might reflect a shared mathematical skill component. Importantly, after these modifications, the **factor loadings did not change markedly** (S1–S9 remained high), which means the underlying structure (one strong factor) is intact, but we've simply accounted for some secondary co-variation. Such an approach is common in measurement modeling to achieve goodness-of-fit while retaining a unidimensional construct, and is justified when error correlations have logical reasons (Byrne, 2010; Kline, 2016).

In summary, the CFA results (especially the modified model) confirm that our instrument measures a **single latent construct of scientific literacy** with high coherence. The high loadings and high *Composite Reliability* attest that the instrument items work well together to capture the intended skills. The initial misfit and subsequent modifications underscore that while one factor is sufficient to explain the lion's share of variance, minor dependencies between items exist (as one would expect in a test where some items share similar formats or knowledge bases). After accounting for these, the instrument's structure is sound.

4.3. Reliability and Item Statistics

Beyond validity, the instrument demonstrated strong reliability. As noted, the ***Composite Reliability*** was 0,947, indicating that the set of indicators is very consistent in measuring the construct. For comparison, *Cronbach's alpha coefficient* for the final test (9 indicators, some averaged from 15 items) was calculated to be 0,91, which falls in the "excellent" range for internal consistency. This suggests that students who scored high on one part of the test tended to score high on other parts, and vice versa, reflecting the unidimensional nature of the test.

The **item-level statistics** were also satisfactory. The difficulty level of the multiple-choice items ranged roughly from 0,30 to 0,80 (proportion of students answering correctly). This spread indicates the test included some relatively easy items and some challenging ones, providing a range that is useful to distinguish different ability levels. The average item difficulty was around 0,55, which is close to ideal for maximizing discrimination in a norm-referenced test. The **item discrimination indices** (point–biserial correlations between item score and total score) were all positive and mostly above 0,40, which is considered very good. No item showed a negative or low discrimination that would signal a problem (e.g., ambiguity or miskeying).

For instance, the item corresponding to indicator S8 (interpreting statistical data) was answered correctly by a relatively high proportion of students (it turned out to be one of the easier items) and also had a high discrimination – likely reflecting that this key item was a good differentiator (many students got it right, especially those with higher overall scores). On the other hand, the item related to creating a graph (S4) was among the more difficult ones (fewer students answered correctly), which might explain its lower factor loading as well. However, that item still had a decent discrimination index (~0.35 point–biserial), meaning it wasn't out of line, but its unique challenge might have made it less correlated with the total score than other items. These observations align with the *CFA* findings and further substantiate the reliability of the instrument.

In addition to internal consistency, we considered **practical reliability** – the stability of results if used in a classroom setting. While test-retest reliability was not formally measured in this study (due to the intervention between pre- and post-test, which changed true scores), the high internal consistency and the strong validation results suggest that the instrument would yield stable and consistent results for a given group of students.

5. Discussion

This study set out to develop and validate a culturally grounded, multidimensional scientific literacy instrument for Indonesian junior high students, integrating nine core indicators, conceptual understanding; scientific argumentation; source evaluation; graph interpretation; quantitative problem solving; data literacy; statistical reasoning; evidence synthesis; and critical reflection, through 15 context-rich multiple-choice items. Content validation by three experts yielded *Aiken's V* coefficients between 0,73 and 0,87 (mean ≈ 0,80), indicating strong consensus on item relevance and clarity.

Classical item analysis confirmed acceptable difficulty (0,31–0,69) and discrimination (point–biserial > 0,30) indices . *Confirmatory factor analysis* (*CFA*) supported a unidimensional structure after modeling three error covariances among closely related indicators, with all standardized loadings ≥ 0,70 (λS8 = 0,99; λS4 = 0,70), *Average Variance Extracted* (*AVE*) = 0,67, and *Composite Reliability* (*CR*) = 0,95, corroborating convergent validity and internal consistency (Cronbach's α ≈ 0,91). These structural relationships are illustrated in Figure 1, which presents the initial CFA model, and in Figure 2, which shows the improved model after applying theoretically justified residual covariances.

The exceptionally high loading for the statistical interpretation indicator (S8) underscores the centrality of *quantitative reasoning* within junior high scientific literacy. Students' ability to interpret statistical data appears to serve as a bellwether for broader scientific reasoning, aligning with research that situates quantitative literacy as foundational for *evidence-based conclusions* (Bauer, 2009; Gormally et al., 2012). The near-perfect loading (λ = 0,99) suggests that mastery of statistical interpretation is nearly synonymous with overall literacy in our sample, reflecting a developmental stage where quantitative tasks strongly differentiate students' competence.

Conversely, the somewhat lower, but still acceptable, loading for graph interpretation (S4; λ = 0,70) reveals a relative gap in data visualization skills, echoing prior findings that graph comprehension presents unique challenges for adolescents. This differential pattern indicates that while students may grasp conceptual and argumentative elements, proficiency in constructing and interpreting graphical representations may require targeted instructional support.

Earlier localized instruments in Indonesia achieved respectable validity and reliability but remained domain-specific. Rusilowati (2016) reported *V* = 0,74 and α = 0,74 for a cycle-theme biology tool, Atta et al. (2020) achieved *V* = 0,77 and *a* = 0,85 for tornado physics items, and Mulyana & Desnita (2023a) obtained *V* scores of 0,81–0,86 with >83% practicality. Pratiwi (2022) and Helendra (2021) likewise demonstrated high content validity (93% and 87,23%, respectively) but confined their focus to biology topics. In contrast, concept inventories (e.g., (Anderson et al., 2002; Smith et al., 2008) target misconceptions narrowly, and TOSLS (Gormally et al., 2012; Quitadamo et al., 2008) addresses undergraduates' *quantitative reasoning* but lacks relevance for younger learners. Globally, Bybee et al. (2009) and Coppi et al. (2023) introduced the ALCE for Portuguese students with robust *CFA* evidence, and PISA-based tool employed Rasch and Bookmarking methods.

While these represent methodological advances, they either span older grades or omit cultural contextualization. Our instrument uniquely merges comprehensive cognitive and quantitative domains with weighted multiple-choice scoring and local scenarios, filling a notable gap in junior high assessment.

The unidimensional factor structure aligns with PISA's practice of reporting a single science literacy score, despite assessing multiple competencies (OECD, 2015). This supports theoretical positions that, for secondary-level cognitive skills, diverse domains of scientific literacy coalesce into a singular latent ability (AAAS, 2011; DeBoer, 2000; National Research Council, 1999). Our findings contribute to debates regarding the dimensionality of scientific literacy by demonstrating that even when capturing critical evaluation, *quantitative reasoning*, and argumentation, a coherent, unified construct emerges when measured via context-rich, rigorously validated items (Pimentel-Velázquez, 2025). Moreover, the necessity to model three error covariances among related indicators reflects local dependencies akin to a bifactorial nuance, an insight that may guide the refinement of theoretical models to acknowledge minor sub-dimensions without fragmenting the overarching construct.

The instrument's high reliability and weighted scoring enable fine-grained diagnosis of student competencies (Pellegrino & Quellmalz, 2010; Schraw, 2010). Educators can use pre-/post-test designs to evaluate interventions, pinScore -ing precisely which indicators, such as graph interpretation or source evaluation, require reinforcement. Embedding items in familiar contexts (cassava cultivation, flood management) enhances authenticity, supports ethnoscience integration, and aligns with recommendations to localize science teaching (Bybee & McCrae, 2011; Kutner et al., 2002; Nuraini & Waluyo, 2021). This approach not only measures literacy but also reinforces relevance, potentially improving student engagement and motivation.

The adoption of weighted multiple-choice scoring distinguishes this instrument from traditional dichotomous formats. By allocating partial credit to defensible distractors and full credit to correct answers, the instrument captures intermediate levels of understanding, thereby increasing score reliability and validity without compromising efficiency. This nuanced scoring enhances diagnostic precision, allowing educators to differentiate students who possess fragmentary conceptual insights from those demonstrating comprehensive mastery.

Several limitations temper the generalizability of these findings. First, the sample was restricted to 11 schools in Central Sulawesi; cross-validation with students from other Indonesian regions or cultural backgrounds is essential to confirm external validity and detect potential cultural biases (Bachman et al., 1995; Fornell & Larcker, 1981). Second, the exclusive use of multiple-choice items, while practical, may not fully capture open-ended reasoning; future iterations could integrate constructed-response tasks to triangulate cognitive processes (Crocker et al., 2008; Hair et al., 2010).

Third, predictive validity remains untested: correlating instrument scores with external criteria (e.g., national exam performance, teacher evaluations) would strengthen claims of criterion-related validity). Finally, discriminant validity was not explicitly examined; future research should demonstrate that the instrument measures scientific literacy distinctly from reading or general mathematical ability by conducting multitrait-multimethod analyses.

Building on this work, researchers should

1) develop alternative test forms with equivalent items to mitigate practice effects;

2) explore computer-based interactive assessments, including simulations, to capture dynamic scientific reasoning;

3) perform differential item functioning (DIF) analyses to ensure fairness across gender and socioeconomic groups (Angoff, 1984; Melser et al., 2020); and

4) implement longitudinal studies to track scientific literacy development over time and identify effective pedagogies.

6. Conclusions

This study presents a rigorously validated, multidimensional scientific literacy instrument tailored to Indonesian junior high contexts. By harmonizing global frameworks with local scenarios, employing dual validation methodologies (*Aiken's V* and CFA), and introducing weighted multiple-choice scoring. The instrument offers educators and researchers a robust tool for diagnosing, monitoring, and enhancing critical thinking and data literacy. As education systems worldwide strive to cultivate 21st-century competencies, this culturally grounded, psychometrically sound assessment model provides a template for integrated, contextually relevant measurement of scientific literacy that can inform curriculum design, instructional strategies, and policy decisions.

6. References

AAAS. (2011). *Vision & change in undergraduate biology education*. American Association for the Advancement of Science. http://visionandchange.org/

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, *45*(1), 131–142. https://doi.org/10.1177/0013164485451012

Alsubait, T., Parsia, B., & Sattler, U. (2013). A similarity-based theory of controlling MCQ difficulty. *2013 Second International Conference on E-Learning and E-Technologies in Education (ICEEE)*, 283-288. https://doi.org/10.1109/ICeLeTE.2013.6644389

Anderson, D. L., Fisher, K. M., & Norman, G. J. (2002). Development and evaluation of the conceptual inventory of natural selection. *Journal of Research in Science Teaching, 39*(10), 952-978. https://doi.org/10.1002/tea.10053

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

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Educational Testing Service.

Atta, H. B., Vlorensius, A., Aras, I., & Ikhsanudin. (2020). Developing an instrument for students' scientific literacy. *Journal of Physics: Conference Series, 1422*(1), 012-019. https://doi.org/10.1088/1742-6596/1422/1/012019

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*(2), 238–257. https://doi.org/10.1177/026553229501200206

Bauer, M. W. (2009). The evolution of public understanding of science: Discourse and comparative evidence. *Science, Technology and Society, 14*(2), 221-240. https://doi.org/10.1177/097172180901400202

Bauer, M. W., Allum, N., & Miller, S. (2007). What can we learn from 25 years of PUS survey research? Liberating and expanding the agenda. *Public Understanding of Science, 16*(1), 79-95. https://doi.org/10.1177/0963662506071287

Bybee, R., & McCrae, B. (2011). Scientific literacy and student attitudes: Perspectives from PISA 2006 science. *International Journal of Science Education, 33*(1), 7-26. https://doi.org/10.1080/09500693.2010.518644

Bybee, R., McCrae, B., & Laurie, R. (2009). PISA 2006: An assessment of scientific literacy. *Journal of Research in Science Teaching, 46*(8), 865-883. https://doi.org/10.1002/tea.20333

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (2nd ed.). Lawrence Erlbaum Associates. https://psycnet.apa.org/record/2009-14446-000

Chevalier, C. D., Ashley, D. C., & Rushin, J. W. (2007). Acquisition and retention of quantitative communication skills in an undergraduate biology curriculum: Long-term retention results. *Journal of College Science Teaching, 39*(2000), 64-71. https://www.jstor.org/stable/42993821

Colon-Berlingeri, M., & Burrowes, P. A. (2011). Teaching biology through statistics: Application of statistical methods in genetics and zoology courses. *CBE – Life Sciences Education, 10*(3), 259–267. https://doi.org/10.1187/cbe.10-11-0137

Coppi, M., Fialho, I., & Cid, M. (2023). Developing a scientific literacy assessment instrument for Portuguese 3rd cycle students. *Education Sciences, 13*(9), 941. https://doi.org/10.3390/educsci13090941

Crocker, L., Algina, J., Staudt, M., Mercurio, S., Hintz, K., & Walker, R. A. (2008). *Introduction to classical and modern test theory*. University Press.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39-50. https://doi.org/10.1177/002224378101800104

Garvin-Doxas, K., & Klymkowsky, M. W. (2008). Understanding randomness and its impact on student learning: Lessons learned from building the Biology Concept Inventory (BCI). *CBE – Life Sciences Education, 7*(2), 227-233. https://doi.org/10.1187/cbe.07-08-0063

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

Gottlieb, M., Bailitz, J., Fix, M., Shappell, E., & Wagner, M. J. (2023). Educator's blueprint: A how-to guide for developing high-quality multiple-choice questions. *AEM Education and Training, 7*(1), e10836. https://doi.org/10.1002/aet2.10836

Hager, P., & Beckett, D. (2020). The emergence of complexity: Rethinking education as a social science. *International Journal of Lifelong Education, 39*(5-6), 453-467. https://doi.org/10.1080/02601370.2020.1860558

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate statistical analysis* (7th ed.). Pearson.

Helendra, H., & Sari, D. R. (2021). Pengembangan instrumen asesmen berbasis literasi sains tentang materi sistem ekskresi dan sistem pernapasan. *Jurnal Ilmiah Pendidikan Profesi Guru, 4*(1), 17-25. https://doi.org/10.23887/jippg.v4i1.29860

Istyadji, M., & Sauqina. (2023). Conception of scientific literacy in the development of scientific literacy assessment tools: A systematic theoretical review. *Journal of Turkish Science Education, 20*(2), 281-308. https://doi.org/10.36681/tused.2023.016

Jia, B., He, D., & Zhu, Z. (2020). Quality and feature of multiple-choice questions in education. *Problems of Education in the 21st Century, 78*(4), 576-594. https://doi.org/10.33225/pec/20.78.576

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8*(2), 147-154. https://doi.org/10.1177/014662168400800202

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). The Guilford Press.

Kutner, M., Greenberg, E., Jin, Y., & Christine. (2002). The health literacy of America's adults: Results from the 2003 National Assessment of Adult Literacy. *Pharmacotherapy, 22*(5), 282-302. https://doi.org/10.1592/phco.22.5.282.33191

Levy, E. H. (2010). *Science in the liberal arts and sciences*. In J. Meinwald & J. G. Hildebrand (Eds.), *Science and the educated American: A core component of liberal education* (41-56). American Academy of Arts and Sciences. https://www.amacad.org/content/publications/publication.aspx?d=333

Maienschein, J. (1998). Scientific literacy. *Science, 281*(5379), 917. https://acortar.link/oUUDfp

Melser, M. C., Steiner-Hofbauer, V., Lilaj, B., Agis, H., Knaus, A., & Holzinger, A. (2020). Knowledge, application, and how about competence? Qualitative assessment of multiple-choice questions for dental students. *Medical Education Online, 25*(1). https://doi.org/10.1080/10872981.2020.1714199

Marsteller, P., de Pillis, L., Findley, A., Joplin, K., Pelesko, J., Nelson, K., Thompson, K., Usher, D., & Watkins, J. (2010). Toward integration: From quantitative biology to mathbio-biomath? *CBE – Life Sciences Education, 9*(3), 165-171. https://doi.org/10.1187/cbe.10-03-0053

Mesci, G., Erdas-Kartal, E., & Yeşildağ-Hasançebi, F. (2025). The effect of argumentation-based inquiry on preservice teachers' understanding of the nature of science, inquiry, and scientific literacy. *Teaching and Teacher Education*, 162, 105064. https://doi.org/10.1016/j.tate.2025.105064

Moreno, R., Martínez, R., & Muñiz, J. (2015). Guidelines based on validity criteria for the development of multiple-choice items. *Psicothema, 27*(4), 388-394. https://doi.org/10.7334/psicothema2015.110

Mulyana, V., & Desnita, D. (2023a). Empirical validity and reliability of the scientific literacy assessment instrument based on the Tornado Physics Enrichment Book. *Jurnal Penelitian Pendidikan IPA*, *9*(5), 3961-3967. https://doi.org/10.29303/jppipa.v9i5.3290

Mulyana, V., & Desnita, D. (2023b). Validity and practicality of scientific literacy assessment instruments based on Tornado Physics Enrichment Book. *Jurnal Ilmiah Pendidikan Fisika Al-Biruni*, *12*(1), 11. https://doi.org/10.24042/jipfalbiruni.v12i1.15934

Nasution, I. B., Liliawati, W., & Hasanah, L. (2019). Development of scientific literacy instruments based on PISA framework for high school students on global warming topic. *Journal of Physics: Conference Series*, 1157, 032063. https://doi.org/10.1088/1742-6596/1157/3/032063

National Committee on Science Education Standards. (1996). *National Science Education Standards* (Vol. 26, No. 4). http://eprints.uanl.mx/5481/1/1020149995.PDF

National Research Council. (1999). *Being fluent with information technology*. National Academy Press.

Nuraini, N., & Waluyo, E. (2021). Pengembangan desain instruksional model project-based learning terintegrasi keterampilan proses sains untuk meningkatkan literasi sains. *Jurnal IPA & Pembelajaran IPA, 5*(1), 101-111. https://doi.org/10.24815/jipi.v5i1.20145

OECD. (2019). *PISA 2018 results: Combined executive summaries* (Vols. I–III). Organization for Economic Co-operation and Development. https://doi.org/10.1787/g222d18af-en

OECD. (2022). *PISA 2022 results (Volume I): The state of learning and equity in education*. Organization for Economic Co-operation and Development. https://acortar.link/2N4P75

Pellegrino, J. W., & Quellmalz, E. S. (2010). Perspectives on the integration of technology and assessment. *Journal of Research on Technology in Education, 43*(2), 119-134. https://doi.org/10.1080/15391523.2010.10782565

Pimentel-Velázquez, C. (2025). CLIL and materials for learning: A rubric for analysis. *European Public & Social Innovation Review*, 10, 1–19. https://doi.org/10.31637/EPSIR-2025-1564

Pratiwi, S. A., Syamsiah, S., & Bahri, A. (2022). Pengembangan instrumen penilaian kemampuan literasi sains pada pembelajaran biologi. *Biology Teaching and Learning, 5*(1), 1-10. https://doi.org/10.35580/btl.v5i1.33834

Quitadamo, I. J., Faiola, C. L., Johnson, J. E., & Kurtz, M. J. (2008). Community-based inquiry improves critical thinking in general education biology. *CBE – Life Sciences Education, 7*(3), 327–337. https://doi.org/10.1187/cbe.07-11-0097

Rusilowati, A. (2016). Pengembangan bahan ajar berbasis literasi sains materi fluida statis. *UPEJ Unnes Physics Education Journal, 5*(3), 25-31. https://doi.org/10.15294/upej.v5i3.13726

Rutherford, J., & Ahlgren, A. (1990). *Science for all Americans*. Oxford University Press.

Santiago, B. J., Ramírez, J. M. O., Rodríguez-Reséndiz, J., Dector, A., García, R. G., González-Durán, J. E. E., & Sánchez, F. F. (2020). Learning management system-based evaluation to determine academic efficiency performance. *Sustainability, 12*(10), 4256. https://doi.org/10.3390/su12104256

Schraw, G. (2010). Measuring self-regulation in computer-based learning environments. *Educational Psychologist, 45*(4), 258-266. https://doi.org/10.1080/00461520.2010.515936

Shi, J., Wood, W. B., Martin, J. M., Guild, N. A., Vicens, Q., & Knight, J. K. (2010). A diagnostic assessment for introductory molecular and cell biology. *CBE – Life Sciences Education, 9*(4), 453-461. https://doi.org/10.1187/cbe.10-04-0055

Smith, M. K., Wood, W. B., & Knight, J. K. (2008). The genetics concept assessment: A new concept inventory for gauging student understanding of genetics. *CBE Life Sciences Education, 7*(4), 422-430. https://doi.org/10.1187/cbe.08-08-0045

Tsui, C. Y., & Treagust, D. (2010). Evaluating secondary students' scientific reasoning in genetics using a two-tier diagnostic instrument. *International Journal of Science Education, 32*(8), 1073-1098. https://doi.org/10.1080/09500690902951429

CONTRIBUCIONES DE AUTORES/AS, FINANCIACIÓN Y AGRADECIMIENTOS

Author Contributions:

Conceptualization: Tadeko, Nurgan; **Methodology**: Tadeko, Nurgan; **Instrument Development**: Tadeko, Nurgan; **Data Collection**: Tadeko, Nurgan; **Formal and Statistical Analysis**: Tadeko, Nurgan; **Data Curation**: Tadeko, Nurgan; **Original Draft Writing**: Tadeko, Nurgan; **Review & Editing**: Rosana, Dadan; Purwastuti, Lusila Andriani; **Visualization**: Tadeko, Nurgan; **Supervision and Academic Guidance**: Rosana, Dadan; Purwastuti, Lusila Andriani; **Project Administration**: Tadeko, Nurgan. All authors have read and approved the final version of the manuscript: Tadeko, Nurgan; Rosana, Dadan; Purwastuti, Lusila Andriani.

Funding: This study did not receive external funding.

Acknowledgments: The authors thank the principals, teachers, and administrative staff of the eleven secondary schools for facilitating access and application of the instrument. They also express their gratitude to the 408 participating students. Furthermore, they acknowledge the valuable contribution of the experts who reviewed content validity, thereby improving the final instrument design. Finally, they extend their thanks to the Educational Assessment Laboratory team at Tadulako University for their support in data processing and statistical review.

Conflict of Interest: The authors declare no conflicts of interest.

AUTHOR/S:

Nurgan Tadeko
Yogyakarta State University, Indonesia.

Nurgan Tadeko is a student at Yogyakarta State University, as well as a researcher and lecturer in the field of science education. He specializes in the development and validation of instruments for measuring scientific literacy in multicultural contexts, employing *confirmatory factor analysis* and advanced statistical modeling techniques for the empirical validation of integrated science learning media. At Tadulako University in Indonesia, he teaches quantitative methodology courses, coordinates multidisciplinary teams with students and cultural tutors, and conducts workshops on SEM/*AMOS* analysis and the development of learning media.

nurgantadeko.2017@student.uny.ac.id

H Index: 2
Orcid ID: https://orcid.org/0000-0003-3805-9078
Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=57220037116
Google Scholar: https://scholar.google.com/citations?user=mc-JPbsAAAAJ&hl=id

Dadan Rosana
Yogyakarta State University, Indonesia.

Dadan Rosana is a professor in the Department of Physical Education at the Faculty of Mathematics and Natural Sciences, Yogyakarta State University, Indonesia. Since 2023, she has served as Dean of the Faculty of Mathematics and Natural Sciences at Yogyakarta State University. She holds a Bachelor of Science from Bandung Teachers' Training Institute (IKIP Bandung), a Master of Science from Bandung Institute of Technology, and a Ph.D. in Educational Measurement and Evaluation from Yogyakarta State University. His research interests include creativity and higher-order thinking skills in science education.

danrosana@uny.ac.id

H Index: 7
Orcid ID: https://orcid.org/0000-0003-4987-7420
Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=57195052389
Google Scholar: https://scholar.google.com/citations?user=OqdW_44AAAAJ&hl=id

Lusila Andriani Purwastuti
Yogyakarta State University, Indonesia

Lusila Andriani Purwastuti has served as an Associate Professor at the Faculty of Education, Yogyakarta State University since 2016. She earned her Ph.D. in Social Pedagogy from Gadjah Mada University in 2015. Her research focuses on inclusive curriculum design, formative assessment of transversal competencies, and critical pedagogy in multicultural settings. She facilitates workshops on competency-based assessment grounded in sociocultural theory and leads international collaborations on social justice and educational equity.

lusila_ap@uny.ac.id

H Index: 3
Orcid ID: https://orcid.org/0000-0003-3385-8700
Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=57219331459
Google Scholar: https://scholar.google.com/citations?user=C5qEtIUAAAAJ&hl=id