International Journal of Instruction *January 2027 ● Vol.20, No.1*
e-ISSN: 1308-1470 ● www.e-iji.net *p-ISSN: 1694-609X*
pp. 673-690

Article submission code: 20260407043925
Received: 07/04/2026 Revision: 11/08/2026
Accepted: 19/08/2026 OnlineFirst: 01/0/2026

Constructing and Pilot-Testing a Rubric for Assessing Lower Secondary Students' Decision-Making Skills in a Socioscientific Issues Context

Suci Indah Putri
Graduated School of Humanities and Social Sciences, Hiroshima University, Japan,
d245910@hiroshima-u.ac.jp

Decision-making is an essential competency for scientific literacy, yet existing rubrics capture only a limited range of its cognitive components. This study constructed and pilot-tested an eight-component rubric to assess lower secondary students' decision-making in a socioscientific issue context using a palm-oil biodiesel expansion scenario with 48 eighth-grade students in Indonesia. The components included identifying problems, collecting scientific data and evidence, considering multiple perspectives, generating options, developing criteria, assessing pros and cons, selecting and justifying decisions, and evaluating decisions. The rubric was developed through four phases. Content validity evidence for the assessment scenario was obtained through expert review and inter-rater reliability evidence through two scoring rounds. Quadratic-weighted kappa was reported in Round 1 (0.718–0.959) and Round 2 (0.810-0.946) after consensus-based refinement, while exact agreement improved for seven of eight components. Within this pilot sample, students performed relatively better in identifying problems and showed difficulty in considering multiple perspectives and evaluating decisions, suggesting potential targets for instruction. The rubric is intended to offer teachers component-level diagnostic information to support targeted decision-making instruction. Because the study used a small sample and a single scenario, further testing with larger and more diverse samples, multiple SSI contexts, and independent raters is needed to establish broader validity evidence.

Keywords: decision-making skills, assessment rubric, iterative refinement, socioscientific issues, inter-rater reliability

INTRODUCTION

Decision-making is one of the essential skills that students must develop through science learning. The PISA 2025 Science Framework (OECD, 2025) identifies the ability to research, evaluate, and use scientific information for decision-making and action as an important characteristic of scientific literacy. Science classrooms are expected to develop students' decision-making skills to make informed and rational decisions. This has also been emphasized by several international organizations (National Research Council, 2012; OECD, 2022; The European Commission, 2019),

Citation: Putri, S. I. (2027). Constructing and pilot-testing a rubric for assessing lower secondary students' decision-making skills in a socioscientific issues context. *International Journal of Instruction*, *20*(1), 673-690.

which highlight the critical importance of developing analytical, rational, and evidence-based decision-making competencies in science education.

Decision-making is a skill that consists of several cognitive components, such as identifying problems, establishing criteria, generating possible alternatives, and analyzing potential solutions (Dauer et al., 2025; Jimenez et al., 2024). The quality of decision-making is not only determined by the decision itself but also by the series of processes through which the decision is made, including the use of optimization strategies (Papadouris, 2012), systematic reasoning (Gresch et al., 2013), and logical structure when constructing the decision (Alred & Dauer, 2020; Dauer et al., 2022; Mehl et al., 2020). Moreover, decision-making is a complex process, which is not merely choosing from existing options but also resolving problems or conflicts among stakeholders (Ben-Horin et al., 2023; Sakamoto et al., 2021). Therefore, assessing decision-making skills requires special attention to each component that reflects the cognitive competencies involved in the process.

However, the cognitive components involved in decision-making do not emerge spontaneously. Students need to be confronted with real-world situations that are ambiguous, complex, and involve conflicting values. Socioscientific issues (SSI) provide this kind of problem context, placing students in dilemmas that require judgment under uncertainty rather than the application of a single correct procedure (Ben-Horin et al., 2023; Dauer et al., 2025; Hsu et al., 2022; Lubis et al., 2022). This ill-structured nature aligns with Vision III scientific literacy, which positions science within its social context rather than treating it solely as a body of knowledge, and defines scientific literacy as the capacity to understand and apply scientific practices to societal issues (Hsu et al., 2022; Siarova et al., 2019; Sjöström, 2025; Valladares, 2021). By engaging with SSI topics such as climate change, biodiversity, or biotechnology, students encounter problems that require weighing evidence and perspectives rather than retrieving a predetermined answer, making SSI a productive context for developing and assessing higher-order thinking skills such as decision-making (Arifin et al., 2025; Nahum et al., 2010; Villarín & Fowler, 2019). Because this process is inherently multidimensional, the instruments used to evaluate it must capture its full complexity rather than isolated components.

Previous studies have conceptualized decision-making through varying and partially overlapping sets of components. Some focus on problem identification, generating alternatives, and selecting solutions (Jimenez et al., 2024; Luan et al., 2022; Papadouris, 2012). Other frameworks emphasize specific sub-skills, such as criteria construction and outcome evaluation (Lee et al., 2019; Sakschewski et al., 2014), the selection phase using optimization or weighted value models (Mettas, 2011; Papadouris, 2012), or argumentative structure (Mehl et al., 2020). Limiting decision-making to a handful of components overlooks its recursive and multidimensional nature (Jimenez et al., 2024; Zhang & Hsu, 2021).

Putri (2026) addressed the diverse conceptualizations of decision-making components by conducting a systematic literature review that synthesizes 20 studies published between 2010 and 2025 into a unified conceptual framework comprising eight

components of decision-making: identifying problems, collecting scientific data and evidence, considering multiple perspectives, generating options, developing criteria, assessing pros and cons, selecting and justifying decisions, and evaluating decisions. This synthesis, however, remained conceptual; it did not include performance descriptors, scoring levels, or empirical evidence of how the framework functions when applied to actual student responses, and the original study explicitly called for future research to examine the framework in classroom contexts. The present study responds to this call by translating the eight components into an assessment rubric with defined scoring levels, examining its inter-rater reliability, and empirically investigating how lower secondary students perform across each component when engaging with an authentic scenario.

The eight components can be treated as separate for assessment purposes, as each draws on qualitatively distinct cognitive skills. For instance, identifying the problem demands analytical framing, developing criteria draws on value-based reasoning, and evaluating the decision outcome relies on metacognitive reflection rather than the generative thinking used to produce alternatives. At the same time, classical and SSI-specific process models of decision-making conceptualize these stages as sequentially and recursively linked rather than fully independent. Simon's (1960) intelligence-design-choice model explicitly allows the decision-maker to loop back to an earlier stage for further validation, and this iterative structure is echoed in SSI-specific staged models of decision-making (Fang et al., 2019). The operational separation of components in the rubric is intended to improve reliability and interpretability rather than to suggest that decision-making components function independently. Instead, the components represent distinct but interconnected cognitive processes within an iterative decision-making framework.

Developing an assessment rubric is therefore essential for understanding the quality of students' decision-making skills. Without clear assessment criteria, it is difficult for teachers to define what constitutes proficient performance or to provide targeted support for students who struggle with specific components (Levinson et al., 2012; Yacoubian, 2018). Structured rubrics enable educators to pinpoint underdeveloped components and provide empirical evidence of student performance across each dimension of decision-making (Gresch et al., 2013; Liu et al., 2011; Papadouris, 2012).

Several rubrics have been proposed. Each reflects a different conceptualization of the construct depending on its intended purpose. Hsu & Lin (2017) developed a rubric targeting individual cognitive abilities through four components: generating criteria, analyzing alternatives, applying strategies, and evaluating results. The rubric does not explicitly include task understanding and planning, which Zhang & Hsu (2021) identified as essential foundations for subsequent decision-making. Sakamoto et al. (2021) evaluated the overall sophistication of students' decision using a holistic 0-5 scale, emphasizing the resolution of conflicts among stakeholders. This approach evaluates the quality of the final decision but does not distinguish students' performance across individual decision-making components. Zhang & Hsu (2021) assessed collaborative decision-making through five dimensions, including defining the decision space, reasoning from scientific evidence, and socially shared regulation. These

instruments differ not only in the components they assess but in what they treat decision-making to be: a sequence of cognitive operations, a holistic quality of judgment, a socially distributed activity, or an inductively derived set of practices. Among the intruments reviewed here, none combines an eight-component analytic structure with a scored rubric designed for individual, component-level diagnosis of lower-secondary students' written reasoning in SSI contexts.

The rubric developed in this study addresses this gap by operationalizing the eight-component framework synthesized by Putri (2026) into scorable descriptors. While the components themselves derive from prior synthesis, the contribution of the present study lies in translating them into descriptors that discriminate between levels of student performance and can be applied consistently by raters. Assessing each component separately allows student profiles to be located within the decision-making process rather than summarized as a single judgment of decision quality, while recognizing that decision-making remains an interconnected and iterative process.

Building on this conceptual foundation, the present study aims to construct an assessment rubric capable of capturing the eight components of students' decision-making and to gather preliminary evidence of its reliability when applied to authentic student responses. Specifically, this study addresses the following research questions: (1) How can the eight components of decision-making synthesized by Putri (2026) be translated into an assessment rubric with clear performance descriptors? (2) To what extent can consistent agreement between raters be achieved in applying this rubric, and what refinement of descriptors is required to reach it? (3) What preliminary patterns of strength and difficulty emerge across the eight components in this sample of students' responses to a socioscientific issue scenario?

METHOD

Research Design

This study employed an instrument development design to develop and preliminarily evaluate an assessment rubric for assessing the eight components of students' decision-making skills in lower secondary science education using a scenario-based test. The rubric was developed through four iterative phases (Chatterji et al., 2002; Sondergeld & Johnson, 2019): (1) planning construct, (2) developing the initial rubric, (3) qualitative field study and data-driven refinement, and (4) reliability testing and consensus-based refinement. As a pilot study, it focuses on content-related review of the scenario and inter-rater reliability of the rubric. Descriptor refinement occurred iteratively across phases, and each phase is described in the subsections that follow.

Participants

This study involved 48 eighth-grade students from two intact classes at a public junior high school located in West Java, Indonesia. Participants were recruited through convenience sampling, which is appropriate for the pilot stage of instrument development. The assessment was administered in a whole-class setting during regular science lessons under the science teacher's supervision, with each student completing the scenario-based open-ended assessment independently on their own device via Microsoft Forms within a ninety-minute session. This pilot study was conducted with

permission from the participating school and the science teachers. Permission for students' participation was granted at the school level. Students were informed of the study's purpose and voluntary nature before completing the assessment. All responses were anonymized before analysis.

Development of the Scenario-Based Assessment

Students' decision-making skills were assessed using a scenario-based test consisting of a context-based SSI and eight open-ended questions, each designed to elicit one component of the decision-making. The assessment was developed around the socioscientific issue of expanding palm-oil biodiesel production in Indonesia, a topic that is both globally significant and locally relevant. The scenario aligns with lower secondary science curriculum content related to ecosystems, biodiversity, and human-environment interactions. It also presents competing economic and environmental considerations, requiring students to evaluate evidence, consider multiple perspectives, and justify decisions rather than identify a single correct answer.

The scenario and assessment questions underwent qualitative expert review prior to pilot implementation. One science education professor and two science teachers with more than five years of teaching experience in junior high school evaluated the alignment between the assessment prompts and their intended decision-making components to examine the content relevance of the instrument. A small panel of subject-matter experts is commonly used to establish content-related evidence for science instruments before pilot administration (Adnan et al., 2021). Review focused on three aspects: (a) relevance, whether each question appropriately elicited its intended component; (b) clarity, the wording and comprehensibility of the prompts; and (c) contextual appropriateness, whether the scenario and questions were suitable and meaningful for Grade 8 students. Feedback from the reviewers was incorporated to revise the wording of the scenario and several assessment prompts before the pilot study. The revisions primarily involved improving the clarity of the scenario narrative, simplifying ambiguous wording, and refining several assessment prompts to better align with their intended decision-making components. The test instrument blueprint was constructed to map each specific prompt question to its corresponding rubric component in Table 1.

Assessment Rubric Construction Process

The construction of the assessment rubric followed the four-phase development process described above, outlined in the subsections below.

Phase 1 – Planning Construct

The planning phase consisted of two main activities: defining the construct to be assessed and reviewing existing decision-making assessment rubrics to determine whether a suitable instrument already existed. Building on the theoretical gap identified in the Introduction, the eight-component construct (Putri, 2026) was integrated with a four-level performance scale adapted from Zhang & Hsu (2021). Prior studies have also consistently used scenario-based assessments with open-ended questions to evaluate decision-making (Gutierez, 2015; Hsu & Lin, 2017; Lee, 2007), which informed the assessment format adopted for Phase 2.

Phase 2 – Developing Initial Rubric

Using the outcomes of Phase 1, an initial assessment rubric was drafted. Each component was scored on a 0-3 scale, with level descriptors intended to differentiate students' responses according to their completeness, relevance, and quality of reasoning. For example, the initial descriptor for Identify Problems distinguished a general problem statement with partial relevance (Score 2) from one that clearly identified the main problem with specific detail and relevance (Score 3). These descriptors underwent iterative refinement in Phase 3.

An initial assessment rubric was established for the eight decision-making components. Each component was translated into a definition, an assessment focus, and corresponding question prompts to guide the assessment of students' responses. These are presented in Table 1.

Table 1
Description of decision-making components and assessment prompts

Component	Definition	Focus of Assessment	Questions Prompt
Identify Problems	The ability to recognize the core dilemma of issue and understand its surrounding context and scientific basis	Identifying the main dilemma, explaining the scientific background and concepts underlying the issue	"Based on the scenario, what is the main problem related to government's plan to expand biodiesel production from palm oil?"
Collect Scientific Data and Evidence	The ability to search for, select, and interpret relevant information from various sources to support inquiry	Searching for and selecting relevant data independent of provided materials, accurately identifying and interpreting evidence.	If you were asked to advise the government about biodiesel policy, what three types of information or data would you need to collect before making a recommendation?
Consider Multiple Perspectives	The ability to analyze the issue through the viewpoints of various stakeholders involved	Identifying potential stakeholders, anticipating diverse views and conflicting values	Identify some perspectives involved in this issue. Describe briefly how each group might view the expansion of biodiesel production.
Generate Options	The ability to brainstorm and propose courses of action	Generating multiple relevant alternative/solutions, listing concrete possible actions to solve the issue	Suggest possible ways that could promote biodiesel while also protecting the environment.
Develop Criteria	The ability to establish standards or rules by which the potential solutions will be judged	Developing relevant criteria, elaborating on criteria with appropriate reasoning	What criteria would you use to decide which biodiesel policy option is best?
Assess Pros and Cons	The ability to evaluate the consequences of each option by applying the established criteria	Weighing arguments for each option, comparing alternatives, determining trade-offs for each alternative	Please write the advantages (pros) and the disadvantages (cons) of the options you suggest.
Select and Justify Decisions	The ability to make a final choice and support it with a logical argument derived from the analysis	Choosing the best option, providing a justification	Based on the options you created, choose the one you think Indonesia should implement. Explain your choice clearly, using reasons and evidence.
Evaluate Decisions	The ability to reflect on the quality of the decision made	Suggesting improvements or revisions to the decision, suggesting the benchmarks of success decisions.	After implementing the biodiesel policy you recommended, what things would you measure or observe to evaluate whether the policy was successful?

Phase 3 – Qualitative Field Testing and Data-Driven Refinement

The scenario-based assessment was administered to the 48 participants, and their open-ended responses were analyzed qualitatively to examine how well the initial rubric functioned in practice. Before conducting inter-rater reliability testing, the researcher examined the range and variability of actual student responses to identify ambiguities, overlaps, and unclear distinctions between score levels. All 48 responses were read for each of the eight components. For each component, response were scored using initial descriptors, and cases in which the descriptor did not clearly determine a single score were recorded together with the source of ambiguity. Applying the initial rubric to authentic responses revealed several descriptors that did not clearly separate different levels of student performance. Based on this, the rubric underwent data-driven refinement: for example, descriptors that failed to distinguish partial from complete responses were reworded to sharpen the boundaries between levels. This data-driven rubric was then used in the first round of inter-rater scoring (Phase 4).

Phase 4 – Reliability Test and Consensus Refinement

Following the data-driven refinement in Phase 3, the reliability of the assessment rubric was examined through a three-step inter-rater scoring and consensus refinement process. Two raters scored the responses: the researcher who developed the rubric and one science teacher from the participating school, with more than five years of teaching experience. Raters were provided with the scoring guide, rubric descriptors, and scoring instructions and independently evaluated the student responses. No discussion of the rubric, scoring criteria, or individual student responses occurred before the initial round of scoring.

Step 1: Initial Blind Scoring (Round 1). The two raters independently scored all 48 responses using the refined (data-driven) rubric.

Step 2: Raters' Agreement Discussion and Consensus. After the initial scoring, the raters participated in a structured consensus meeting to review discrepancies identified in Round 1. Drawing on both the observed scoring disagreements and the teacher's classroom experience, they examined rubric descriptors that were ambiguous, difficult to interpret, or insufficiently aligned with students' responses. These descriptors were subsequently revised to improve clarity, consistency, and applicability while preserving the underlying decision-making constructs.

Step 3: Independent rescoring (Round 2). Using the consensus-based refinement rubric, the two raters again independently scored all 48 responses. Because the same set of 48 student responses was used for field testing, descriptor refinement, and both rounds of reliability testing, the Round 2 agreement should be interpreted cautiously. The observed improvement in inter-rater agreement may partly reflect increased rater calibration to this specific dataset in addition to improvement in descriptor clarity. Consequently, the findings provide preliminary evidence of rubric's scoring reliability rather than definitive evidence of its generalizability across different raters or student samples.

Data Analysis

Data analysis was conducted in two parts: (1) evaluating the inter-rater reliability of the scoring rubric and (2) profiling students' decision-making performance. Inter-rater reliability was evaluated using weighted Cohen's kappa with linear and quadratic weights for each of the eight decision-making components across both rounds of scoring, selected because the rubric employed an ordinal four-point scale (0-3) that allows partial agreement between adjacent score categories to be recognized (Cohen, 1968). Percent agreement (exact match and within-one-category match) was also calculated for each component to characterize the distribution of rater disagreement, and 95% confidence intervals were computed for all kappa coefficients in JASP 0.96. To profile students' decision-making performance, descriptive statistics were computed based on Rater 1 scoring using the rubric's final version. Students' scores were summarized as means (M) and standard deviations (SD) for each of the eight decision-making components to identify relative strengths and areas requiring further instructional support.

FINDINGS

Development and Refinement of the Rubric

Qualitative analysis of the 48 student responses during field testing revealed ambiguities in the initial rubric's performance descriptors, prompting specific data-driven revisions summarized in Table 2. Seven of the eight components required revision; the pattern was dominated by two issues: a rigid, fixed-count requirement at Score 3 that did not discriminate well across the range of actual student responses, and unresolved ambiguity about what a given proficiency level required. Only "Generate Options" required no substantive revision at this stage.

Table 2
Data-driven refinement: initial rubric to data-driven rubric

Component	Issue Identified	Nature of Revision
Identify Problems	Initial rubric could not distinguish between simply identifying the topic and identifying the specific problem and its impacts	Score 3 was revised to require identification of the main problem, its impacts, and related environmental/sustainability concerns.
Collect Scientific Data and Evidence	Score 3 relied on a fixed number of data types, while the descriptors for Scores 1 and 2 did not clearly distinguish different levels of evidence quality.	Descriptors were revised to emphasize evidence quality and clearer progression across score levels.
Consider Multiple Perspectives	Score 3 relied on a fixed stakeholder count, whereas the descriptors for Scores 1 and 2 lacked clear differentiation between levels of perspective-taking.	Descriptors were revised to remove the fixed-count requirement and establish clearer progression across score levels.
Generate Options	No substantive issue identified at this stage	No change to scoring criteria
Develop Criteria	Initial descriptors combined the number and relevance of criteria without clearly distinguishing performance levels, resulting in inconsistent scoring.	Descriptors were refined by specifying the expected number of relevant criteria at each score level to improve scoring consistency.
Assess Pros and Cons	The initial descriptor was ambiguous regarding whether students should evaluate one option or multiple options.	Descriptors were revised to clearly distinguish evaluation of one option (Score 2) from multiple options (Score 3).
Select and Justify	The initial descriptors emphasized the	Descriptors were revised to clarify

Decisions	presence of justification rather than its quality, resulting in limited differentiation between performance levels.	distinctions in the quality of students' justifications.
Evaluate Decisions	The initial descriptors contained ambiguous terms (e.g., "general idea" and "very vague"), making consistent scoring difficult.	Descriptors were revised to use more specific and measurable criteria, establishing clearer distinctions across score levels.

To illustrate the data-driven refinement process, the *Select and Justify Decisions* component was examined using variation in students' responses. One student proposed *"improving human resources and infrastructure"* without providing a supporting reason, while another recommended, "*using non-forest materials because this could reduce forest loss".* Other responses included several relevant considerations, such as *"forest protection, waste reduction, and the use of cooking oil as fuel",* whereas the most developed response *connected sustainable palm-oil development with environmental, economic, and social consequences.* Because the initial descriptors mainly distinguished the presence or absence of reasoning, they did not adequately capture these differences in justification quality. The descriptors were therefore revised to distinguish inadequate support, relevant but limited reasons, and strong, well-developed justification.

Inter-rater Reliability and Consensus-Based Refinement

Following the first round of blind scoring, the data-driven rubric was further refined through consensus discussions with the teacher rater. The revisions focused on improving practical usability and reducing ambiguity in scoring.

Table 3
Consensus-based refinement: data-driven rubric to consensus-refined rubric

Component	Limitation Identified	Key Refinement
Identify Problems	Criteria were too specific to environmental issues, limiting applicability	Revised wording to focus on the main issue, causes, and impacts regardless of the topic.
Collect Scientific Data and Evidence	Phrases like "various data" or "some data" led to inconsistent scoring.	Added specific counts (eg, '3 or more types') to objectively measure breadth.
Consider Multiple Perspectives	Difficulty distinguishing students who gave many general perspectives vs. fewer detailed ones.	Explicitly required "2 or more perspectives" explained in detail for the highest score.
Generate Options	Ambiguity between "general options" (common sense) and "scientific alternatives".	Differentiated levels based on whether options are "scientifically valid" vs. "general alternatives".
Develop Criteria	No limitation identified; the initial descriptors yielded high inter-rater agreement ($\kappa = 0.917$ linear, 0.959 quadratic in Round 1)	Retained, original descriptors were maintained without modification.
Assess Pros and Cons	"Pros and cons for each option" was hard to measure consistently	Specified the exact number of alternatives and required "scientific evidence" for pros/cons.
Select and Justify Decisions	Terms like "strong decision" or "relevant reasons" were vague.	Revised descriptors to strictly require "scientific justification (data, concept, mechanism)"
Evaluate Decisions	"Specific and measurable plan" was broadly interpreted	Defined "measurable" as having explicit benchmarks and indicators

For example, the response, *"Avoid clearing new forests and use waste cooking oil as fuel to protect the environment and reduce waste,"* received Scores 2 and 1 from the two raters. One rater considered the response to contain sufficient reasoning for Score 2, although the justification lacked detail; the other judged the reasoning as too vague and inadequate for that level. Disagreement indicated that the initial descriptors did not clearly distinguish limited reasoning from inadequate justification. The descriptors were therefore revised so that Score 2 represented a decision supported by general reasoning but lacking scientific support, whereas Score 1 represented a decision supported only by vague or inadequate justification.

Reliability Outcomes

Inter-rater reliability was examined across all eight components in two scoring rounds: Round 1 using the data-driven rubric and Round 2 using consensus-refined rubric. Table 4 reports linear-and quadratic-weighted Cohen's kappa with 95% confidence intervals and exact and within-one-category agreement for both rounds.

Table 4
Inter-rater reliability across two scoring rounds

Component	Linear-Weighted κ (95% CI)		Quadratic-Weighted κ (95% CI)		% Exact Agreement		% Within-1-Category	
	R1	R2	R1	R2	R1	R2	R1	R2
Identify Problems	0.725 (0.597 - 0.853)	0.756 (0.630- 0.883)	0.860 (0.784– 0.936)	0.858 (0.777– 0.939)	66.7	75.0	100	100
Collect Scientific Data and Evidence	0.818 (0.694- 0.941)	0.893 (0.811- 0.976)	0.895 (0.805– 0.985)	0.943 (0.898– 0.989)	81.3	87.5	97.9	100
Consider Multiple Perspectives	0.853 (0.764- 0.941)	0.848 (0.734- 0.961)	0.931 (0.886– 0.976)	0.890 (0.791– 0.989)	81.3	85.4	100	95.8
Generate Options	0.661 (0.508- 0.814)	0.767 (0.637- 0.897)	0.774 (0.637– 0.910)	0.862 (0.777– 0.948)	66.7	77.1	95.8	100
Develop Criteria	0.917 (0.837- 0.997)	0.896 (0.816- 0.975)	0.959 (0.918– 1.000)	0.946 (0.903– 0.989)	91.7	87.5	100	100
Assess Pros and Cons	0.699 (0.535- 0.863)	0.748 (0.608- 0.887)	0.833 (0.710– 0.956)	0.839 (0.740– 0.938)	72.9	77.1	97.9	100
Select and Justify Decisions	0.608 (0.429- 0.786)	0.769 (0.627- 0.910)	0.718 (0.547– 0.890)	0.810 (0.668– 0.952)	66.7	81.3	95.8	95.8
Evaluate Decisions	0.646 (0.460- 0.836)	0.828 (0.735- 0.921)	0.772 (0.604– 0.939)	0.915 (0.864– 0.965)	77.1	79.2	95.8	100

Note: R1 = Round 1, data-driven rubric; R2 = Round 2, consensus-refined rubric

Quadratic-weighted values exceeded linear-weighted values for every component. This ordering follows from the weighting schemes themselves, since quadratic weights penalise one-category disagreements less heavily than linear weights, and the linear-weighted coefficients therefore represent the more conservative estimate. Evidence that

disagreements were in fact concentrated in adjacent categories comes instead from the within-one-category agreement, which ranged from 95.8% to 100% across both rounds.

Overview of Students' Decision-Making Performance

Descriptive analysis results for the 48 students are presented in Table 5. Within this sample, students demonstrated the highest proficiency in *Identify Problems* (M = 1.79), with 62.5% of students achieving a score of 2 or higher. The lowest mean scores were observed in *Evaluate Decisions* (M = 1.15) and *Consider Multiple Perspectives* (M = 1.15). The *Consider Multiple Perspectives* component recorded the highest proportion of students scoring at level 0 (39.6%). For the *Generate Options* and *Assess Pros and Cons* components, the largest proportion of students scored at level 1 (56.3% and 45.8%, respectively).

Table 5
Descriptive statistics of students' decision-making proficiency (N = 48)

Components		Mean (M)	SD	Percentage			
				Score 0	Score 1	Score 2	Score 3
1.	Identify Problems	1.79	0.967	10.4	27.1	35.4	27.1
2.	Collect Scientific Data and Evidence	1.31	1.095	29.2	29.2	22.9	18.8
3.	Consider Multiple Perspectives	1.15	1.130	39.6	22.9	20.8	16.7
4.	Generate Options	1.50	0.945	8.3	56.3	12.5	22.9
5.	Develop Criteria	1.21	1.110	35.4	25.0	22.9	16.7
6.	Assess Pros and Cons	1.21	0.874	20.8	45.8	25.0	8.3
7.	Select and Justify Decisions	1.54	0.944	12.5	39.6	29.2	18.8
8.	Evaluate Decisions	1.15	1.031	31.3	37.5	16.7	14.6

DISCUSSION

This pilot study constructed and refined a component-based rubric for assessing lower secondary students' decision-making in a socioscientific issue context. Two main findings emerged. First, the scoring descriptors required iterative refinement before they could be applied consistently to varied student responses. Second, the component-level results revealed preliminary differences in students' performances, with relatively stronger performance in identifying the problem and difficulty in considering multiple perspectives and evaluating decisions. The findings provide preliminary evidence that the rubric descriptor scores were distributed across all four levels for every component (Table 5), suggesting that the descriptors did not collapse responses into a narrow band; whether they discriminate levels of quality in the intended sense requires evidence beyond rater agreement. The resulting component-level profiles also indicate the rubric's potential to identify students' relative strengths and difficulties within this assessment context.

The two refinement stages moved in opposite directions with respect to numerical requirement. During data-driven refinement, fixed-count requirements were removed from Collect Scientific Data and Evidence and Consider Multiple Perspectives because they distinguished responses by quantity rather than quality. During consensus refinement, however, both raters reported that the resulting count-free descriptors,

phrased as "various data" or "some perspectives", were difficult to apply consistently, and this was reflected in the Round 1 disagreements for these components. Minimum counts were therefore reinstated as a threshold criterion. This sequence illustrates a trade-off between construct representation and scoring consistency: count-based descriptors are easier to apply reliably but risk rewarding breadth over depth. The final descriptors combine a minimum count with a quality requirement at each level to balance these concerns, but whether this balance is appropriate requires evidence beyond rater agreement.

The reliability results indicate that the main scoring challenge concerned distinctions between neighboring score levels. In Round 1, quadratic-weighted kappa ranged from 0.718 to 0.959, and linear-weighted kappa ranged from 0.608 to 0.917. This difference suggests that the raters rarely assigned widely separated scores, although disagreements between adjacent categories remained more visible under the stricter linear weighting. The consensus process therefore focused on clarifying the boundaries between nearby score levels.

Following consensus-refinement, weighted kappa remained at similar levels across all components. Confidence intervals overlapped between rounds for every component, so neither the increases nor the decreases can be attributed to descriptor revision. The slight decline for *Develop Criteria*, whose descriptor was unchanged between rounds, may reflect ordinary variation in rater judgments or score distributions rather than deterioration in the rubric. For *Consider Multiple Perspectives*, exact agreement improved, and weighted kappa decreased slightly because a small number of responses produced larger disagreements when the breadth and depth of perspective-taking indicated different score levels. These component-specific results show that consensus refinement supported scoring consistency in several areas but did not produce uniform improvement across the rubric. The Round 2 results should also not be attributed solely to descriptor quality. The same raters rescored the same student responses after participating in consensus discussions, and increased familiarity with the responses and a shared interpretation of the descriptors may have contributed to the observed agreement.

The forms of evidence examined in this study should be distinguished carefully. Expert review provided content validity evidence concerning the relevance, clarity, and contextual appropriateness of the assessment scenario and prompts. Teacher review and discussion provided preliminary evidence regarding the practical application of the rubric. The two scoring rounds provided inter-rater reliability evidence, while the examination of actual student responses informed the refinement of the descriptors. However, the study did not formally investigate students' response processes through cognitive interviews or think-aloud procedures. It also did not examine the rubric's internal structure or establish construct validity. The findings should therefore be interpreted as pilot evidence supporting rubric development and scoring reliability rather than as comprehensive validation.

Beyond the development and reliability findings, the component-level results provided preliminary diagnostic evidence of students' decision-making performance in the palm-

oil biodiesel scenario. Students showed the highest mean performance in *Identify Problems* and the lowest mean performance in *Consider Multiple Perspectives* and *Evaluate Decisions*. This variation supports the view that decision-making involves related but distinguishable reasoning processes and that relative strength in one component does not necessarily indicate comparable performance in another.

The component profiles also illustrate the potential value of assessing decision-making. A student may recognize the central problem but still experience difficulty interpreting stakeholder positions, comparing consequences, justifying a choice, or proposing a plan to evaluate its outcomes. An overall score alone would provide limited information about these differences, whereas the component-level rubric makes them more visible.

The lower performance in *Consider Multiple Perspectives* may reflect the additional cognitive demands of perspective-taking. Identifying the problem mainly required students to recognize the central conflict and its consequences from the scenario. Perspective-taking required them to identify relevant stakeholders, interpret their interests and concerns, and explain how their positions differed. Several responses named the parties involved without explaining their viewpoints, suggesting that recognizing stakeholders was more accessible than reasoning from their perspectives.

The difficulty in *Evaluate Decisions* may similarly reflect the demands of reflective and forward-looking reasoning. Students were expected not only to identify what should be evaluated but also to propose indicators, criteria, or benchmarks for determining whether the selected decision had been successful. Many responses mentioned broad outcomes, such as environmental improvement or economic benefit, without explaining how these outcomes could be monitored or measured. The findings therefore suggest that some students had difficulty translating a final decision into a concrete evaluation plan and connecting the later evaluative stage with the reasoning developed earlier in the task.

These component-specific patterns suggest hypotheses for SSI instruction that would need to be tested with classroom data. Perspective-taking could be supported through stakeholder maps, contrasting position statements, role-based discussion, and prompts requiring students to explain the evidence, interests, and concerns underlying each viewpoint. Decision evaluation could be scaffolded by asking students to identify expected outcomes, establish measurable indicators, specify benchmarks, and explain when a decision should be reconsidered.

The rubric can support such instruction by helping teachers identify the particular reasoning process requiring development. Rather than simply judging whether a student selected an acceptable option, teachers can determine whether the difficulty lies in collecting evidence, considering perspectives, developing criteria, comparing consequences, justifying the decision, or evaluating its outcomes. This information can guide more targeted feedback and help teachers design activities that address specific weaknesses.

For classroom use, standardized scoring guidance and annotated response examples may be especially important. The present findings suggest that some descriptors were

already interpreted consistently by the two raters, whereas others benefited from discussion and calibration. Although paired scoring may not be practical in routine teaching, benchmark responses and brief rater training could help individual teachers apply the rubric more consistently.

The rubric also contributes to science education by extending assessment beyond students' recall or application of scientific knowledge. It examines how students use evidence, stakeholder perspectives, evaluation criteria, justification, and reflection when responding to a socially and environmentally relevant scientific issue. In this way, the rubric may support scientific literacy by making the reasoning required for informed SSI decision-making more explicit.

Its component-based structure also has potential relevance to Vision III, which connects scientific understanding with participation in social and environmental issues and responsible action. However, the present study assessed students' reasoning about decisions rather than their actual participation, agency, or action. The connection to Vision III should therefore be understood as a potential instructional contribution that requires further empirical examination.

Several limitations should be considered. First, the rubric was pilot-tested with 48 eighth-grade students, so the observed performance profile should be interpreted as preliminary evidence from this sample rather than as representative of lower secondary students more broadly. Second, the assessment used a single palm-oil biodiesel scenario. Because decision-making may depend on topic familiarity, contextual knowledge, and the nature of the stakeholder conflict, the findings do not establish that the rubric functions similarly across SSI contexts. Third, the same responses were used for qualitative refinement, scoring, consensus discussion, rescoring, and for the performance profile reported in Table 5. This design may have increased agreement through rater calibration and familiarity with the dataset. Future studies should test the rubric with larger and more diverse samples, multiple SSI scenarios, and independent raters who were not involved in its development. Further research should also examine students' response processes, the rubric's internal structure, and additional sources of construct validity evidence. Standardized scoring guidelines, annotated examples, and rater-training materials should be developed before broader classroom or large-scale implementation.

CONCLUSION

This pilot study developed and refined an eight-component rubric for assessing lower secondary students' decision-making in a socioscientific issue context. Expert review provided content validity evidence for the assessment scenario, while the two-round scoring process provided preliminary inter-rater reliability evidence for the rubric. Consensus-based refinement was used to clarify ambiguous descriptors and distinctions between adjacent score levels. Scoring agreement remained at similar levels afterwards, although this cannot be attributed to descriptor revision alone, because the same rater rescored the same responses following discussion. Nevertheless, these findings represent preliminary evidence, particularly because the rubric was tested with a small

sample, a single SSI scenario, and the same raters and student responses across both scoring rounds.

Within this assessment context, the rubric differentiated levels of response quality and generated component-level profiles of students' decision-making performance. Students performed relatively better in identifying the problem but showed greater difficulty in considering multiple perspectives and developing measurable plans for evaluating their decisions. These patterns suggest that SSI instruction may need to provide more explicit scaffolding for perspective-taking and reflective evaluation. For science education and educational assessment, the rubric offers a potential diagnostic framework that enables teachers to identify specific strengths and difficulties rather than assigning only a general decision-making score. Future research should test the rubric with larger and more diverse samples, multiple SSI scenarios, and independent raters using standardized scoring guidance and training.

REFERENCES

Adnan, Mulbar, U., Sugiarti, & Bahri, A. (2021). Scientific literacy skills of students: Problem of biology teaching in junior high school in South Sulawesi, Indonesia. In *International Journal of Instruction* (Vol. 14, Number 3, pp. 847–860). Gate Association for Teaching and Education. https://doi.org/10.29333/iji.2021.14349a

Alred, A. R., & Dauer, J. M. (2020). Understanding factors related to undergraduate student decision-making about a complex socio-scientific issue: Mountain lion management. *Eurasia Journal of Mathematics, Science and Technology Education*, *16*(2), 1–21. https://doi.org/10.29333/ejmste/113757

Arifin, Z., Sukarmin, S., & Saputro, S. (2025). Trends and research frontiers in socioscientific issues for sustainable science education: A systematic and bibliometric analysis from 2014 – 2024. *Journal of Pedagogical Research*, *9*(1), 407–434. https://doi.org/10.33902/JPR.202530575

Ben-Horin, H., Kali, Y., & Tal, T. (2023). The Fifth Dimension in Socio-Scientific Reasoning: Promoting Decision-Making about Socio-Scientific Issues in a Community. *Sustainability (Switzerland)*, *15*(12), 1–23. https://doi.org/10.3390/su15129708

Chatterji, M., Sentovich, C., Ferron, J., & Rendina-Gobioff, G. (2002). Using An Iterative Model To Conceptualize, Pilot Test, and Validate Scores From An Instrument Measuring Teacher Readiness For Educational Reforms. *Educational and Psychological Measurement*, *62*(3), 444–465.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220. https://doi.org/10.1037/h0026256

Dauer, J. M., Kirby, C. K., & Sorensen, A. E. (2025). Defining students' socioscientific issues classroom decision-making components and practice proficiencies. *Disciplinary and Interdisciplinary Science Education Research*, *7*(12).

Dauer, J. M., Sorensen, A. E., & Jimenez, P. C. (2022). Using Structured Decision-Making in the Classroom to Promote Information Literacy in the Context of Decision-Making. *Journal of College Science Teaching*, *51*(6), 75–82.

Fang, S. C., Hsu, Y. S., & Lin, S. S. (2019). Conceptualizing Socioscientific Decision Making from a Review of Research in Science Education. *International Journal of Science and Mathematics Education*, *17*(3), 427–448. https://doi.org/10.1007/s10763-018-9890-2

Gresch, H., Hasselhorn, M., & Bögeholz, S. (2013). Training in Decision-making Strategies: An approach to enhance students' competence to deal with socio-scientific issues. *International Journal of Science Education*, *35*(15), 2587–2607. https://doi.org/10.1080/09500693.2011.617789

Gutierez, S. B. (2015). Integrating socio-scientific issues to enhance the bioethical decision-making skills of high school students. *International Education Studies*, *8*(1), 142–151. https://doi.org/10.5539/ies.v8n1p142

Hsu, Y. S., & Lin, S. S. (2017). Prompting students to make socioscientific decisions: embedding metacognitive guidance in an e-learning environment. *International Journal of Science Education*, *39*(7), 964–979. https://doi.org/10.1080/09500693.2017.1312036

Hsu, Y.-S., Tytler, R., & White, P. J. (2022). *Innovative Approaches to Socioscientific Issues and Sustainability Education*.

Jimenez, P. C., Zwickle, A., & Dauer, J. M. (2024). Defining and describing students' socioscientific issues tradeoffs practices. *International Journal of Science Education, Part B: Communication and Public Engagement*, *14*(3), 277–293. https://doi.org/10.1080/21548455.2023.2263608

Lee, Y. C. (2007). Developing decision-making skills for socio-scientific issues. *Journal of Biological Education*, *41*(4), 170–177. https://doi.org/10.1080/00219266.2007.9656093

Lee, Y. C., Grace, M., Rietdijk, W., & Lui, Y. C. (2019). A cross-cultural, cross-age, and cross-gender study of Hong Kong and UK secondary students' decision making about a biological conservation issue. *International Journal of Science Education*, *41*(18), 2696–2715. https://doi.org/10.1080/09500693.2019.1693662

Levinson, R., Kent, P., Pratt, D., Kapadia, R., & Yogui, C. (2012). Risk-based decision making in a scientific issue: A study of teachers discussing a dilemma through a microworld. *Science Education*, *96*(2), 212–233. https://doi.org/10.1002/sce.21003

Liu, S. Y., Lin, C. S., & Tsai, C. C. (2011). College students' scientific epistemological views and thinking patterns in socioscientific decision making. *Science Education*, *95*(3), 497–517. https://doi.org/10.1002/sce.20422

Luan, H., Li, T. L., & Lee, M. H. (2022). High School Students' Environmental Education in Taiwan: Scientific Epistemic Views, Decision-Making Style, and Recycling Intention. *International Journal of Science and Mathematics Education*, *20*(1), 25–44. https://doi.org/10.1007/s10763-020-10136-z

Lubis, S. P. W., Suryadarma, I. G. P., Paidi, P., & Yanto, B. E. (2022). The Effectiveness of Problem-based learning with Local Wisdom oriented to Socio-Scientific Issues. *International Journal of Instruction*, *15*(2), 455–472. https://doi.org/10.29333/iji.2022.15225a

Mehl, C. E., Jin, H., & Llort, K. F. (2020). Student Decision Making in a Scenario-based Investigation of an Ecosystem. *Eurasia Journal of Mathematics, Science and Technology Education*, *16*(1), 1–14. https://doi.org/10.29333/ejmste/112278

Mettas, A. (2011). The development of decision-making skills. *Eurasia Journal of Mathematics, Science and Technology Education*, *7*(1), 63–73. https://doi.org/10.12973/ejmste/75180

Nahum, T. L., Ben-Chaim, D., Azaiza, I., Herskovitz, O., & Zoller, U. (2010). Does stes-oriented science education promote 10th-grade students' decision-making capability? *International Journal of Science Education*, *32*(10), 1315–1336. https://doi.org/10.1080/09500690903042533

National Research Council. (2012). A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*, 1–385. https://doi.org/10.17226/13165

OECD. (2022). *Trends Shaping Education 2022*. OECD Publishing.

OECD. (2025). *PISA 2025 Science Framework*. OECD PISA Framework. https://pisa-framework.oecd.org/science-2025/

Papadouris, N. (2012). Optimization as a reasoning strategy for dealing with socioscientific decision-making situations. *Science Education*, *96*(4), 600–630. https://doi.org/10.1002/sce.21016

Putri, S. I. (2026). Exploring Decision-Making Skills in Science Education: A Systematic Review of Operational Definition and Instructional Trends. *International Journal of STEM Education for Sustainability*, *6*(2), 178–196. https://doi.org/10.52889/ijses.v6i2.873

Sakamoto, M., Yamaguchi, E., Yamamoto, T., & Wakabayashi, K. (2021). An intervention study on students' decision-making towards consensus building on socio-scientific issues. *International Journal of Science Education*, *43*(12), 1965–1983. https://doi.org/10.1080/09500693.2021.1947541

Sakschewski, M., Eggert, S., Schneider, S., & Bögeholz, S. (2014). Students' Socioscientific Reasoning and Decision-making on Energy-related Issues-Development of a measurement instrument. *International Journal of Science Education*, *36*(14), 2291–2313. https://doi.org/10.1080/09500693.2014.920550

Siarova, H., Sternadel, D., & Szőnyi, E. (2019). Science and Scientific Literacy as an Educational Challenge. *EPRS: European Parliamentary Research Service. Belgium*, (March), 1–64.

Simon, H. A. (1960). *The new science of management decision*. Harper & Row.

Sjöström, J. (2025). Vision III of scientific literacy and science education: an alternative vision for science education emphasising the ethico-socio-political and relational-existential. *Studies in Science Education*, *61*(2), 239–274. https://doi.org/10.1080/03057267.2024.2405229

Sondergeld, T. A., & Johnson, C. C. (2019). Development and validation of a 21st Century Skills Assessment: Using an iterative multimethod approach. *School Science and Mathematics*, *119*(6), 312–326. https://doi.org/10.1111/ssm.12355

The European Commission. (2019). Key competences for lifelong learning - Publications Office of the EU. *European Union*, 1–20. https://doi.org/10.2766/291008

Valladares, L. (2021). Scientific Literacy and Social Transformation: Critical Perspectives About Science Participation and Emancipation. In *Science and Education* (Vol. 30, Number 3). Springer Netherlands. https://doi.org/10.1007/s11191-021-00205-2

Villarín, L. J. R., & Fowler, S. R. (2019). Socioscientific issues to promote content knowledge & socioscientific reasoning in puerto rican high school students. *American Biology Teacher*, *81*(5), 328–332. https://doi.org/10.1525/abt.2019.81.5.328

Yacoubian, H. A. (2018). Scientific literacy for democratic decision-making. *International Journal of Science Education*, *40*(3), 308–327. https://doi.org/10.1080/09500693.2017.1420266

Zhang, W. X., & Hsu, Y. S. (2021). The interplay of students' regulation learning and their collective decision-making performance in a SSI context. *International Journal of Science Education*, *43*(11), 1746–1778. https://doi.org/10.1080/09500693.2021.1933250