Journal on Mathematics Education
Volume 13, No. 4, 2022, pp. 587-604

An integrated STEM-based mathematical problem-solving test: Developing and reporting psychometric evidence

Ijtihadi Kamilia Amalina[1], Tibor Vidákovich[2]

[1]Doctoral School of Education, University of Szeged, Hungary
[2]Institute of Education, University of Szeged, Hungary
*Correspondence: ijtihadi.kamilia.amalina@edu.u-szeged.hu

Received: 15 October 2022 | Revised: 3 December 2022 | Accepted: 8 December 2022 | Published Online: 11 December 2022

Abstract

Science, technology, engineering, and mathematics (STEM) problem-solving is necessary to be infused into the classroom. Nevertheless, the criticism of underrepresented mathematics in STEM problem-solving assessment is an issue. In this study, we develop and investigate the psychometric evidence of an integrated STEM-based mathematical problem-solving test. The product of the test was a mathematical essay test that contains three scientific scenarios related to the environment in every middle school grade. The mathematical contents were integrated into engineering-based design using the technology. Three experts filled an assessment sheet to assess content validity, which was analyzed using a content validity index (CVI) and intraclass correlation coefficient (ICC). The result of content validity revealed that overall items were valid and reliable. The construct validity was examined using the Rasch analysis from the data of Grades 7–9 students in Indonesia (n = 286). The construct of all scenarios and prompting items indicated fit with various difficulty levels and acceptable discrimination value. Nevertheless, four prompting items were reported as misfit based on unweighted mean square value. The recommendation for improvement is emphasized in the language clarity aspect. The inter-rater reliability was also declared good. A further study is suggested to provide a computer-based test.

Keywords: Development, Mathematical Problem-Solving, Psychometric Evidence, STEM, Test

How to Cite: Amalina, I. K., & Vidákovich, T. (2022). An integrated STEM-based mathematical problem-solving test: Developing and reporting psychometric evidence. *Journal on Mathematics Education, 13*(4), 587-604. http://doi.org/10.22342/jme.v13i4.pp587-604

Mathematics is a foundation for learning other disciplines and plays a pivotal role in our daily lives (Junpeng et al., 2020; Kesorn et al., 2020; Khairani & Sahari Nordin, 2011). However, graduates encounter difficulties in applying mathematics in real-life situations because practice-oriented mathematics education is isolated from the mathematics curriculum (Jones et al., 2015; Ke & M. Clark, 2020; Shute et al., 2016). This issue causes a shift in mathematics class to problem-solving mathematics in the real-world context (Jones et al., 2015). Problem-solving is the most significant cognitive activity in professional, every day, 21st century, and life-long learning (İncebacak & Ersoy, 2016; Karatas & Baki, 2013). Therefore, problem-solving in mathematics is a core standard of the mathematics curriculum in several countries (Du Toit & Du Toit, 2013; Karatas & Baki, 2013).

Problem-solving in mathematics includes procedure application, concept reasoning, synthesizing, analyzing, accessing information, and interpreting (Karatas & Baki, 2013). Infusing problem-solving in mathematics classes can improve mathematics achievement (Karatas & Baki, 2013). Several researchers have developed, validated, and examined students' problem-solving skills, particularly using closed-ended essay tests (Du Toit & Du Toit, 2013; İncebacak & Ersoy, 2016; Widodo et al., 2021). However, they used well-defined, monodisciplinary, and structured tests, which are fundamentally different from real-life problems (Shute et al., 2016). Hence, they encountered difficulties in both daily demands and international assessment of authentic problem solving (Shanta, 2019; Shute et al., 2016).

Advances in technological innovation and the economic industry globally have impacted the labor market (Kelley & Knowles, 2016; Maass et al., 2019). Professionals who can communicate and use information; solve issues in science, technology, engineering, and mathematics (STEM); adapt; and be innovative in response to changing demands and information are required (Maass et al., 2019). Moreover, global challenges, including social and environmental issues, increase rapidly (Jolly, 2016; Kelley & Knowles, 2016). Hence, the education area needs to improve students' knowledge in interdisciplinary by integrating STEM disciplines (Jolly, 2016; Kelley & Knowles, 2016; Maass et al., 2019). Because STEM is a complex concept, integrating its discipline requires problem solving (Maass et al., 2019).

Over the years, several researchers have developed frameworks related to STEM for formative assessment (Kelley & Knowles, 2016; Maass et al., 2019; Priemer et al., 2020; Wells, 2016). Researchers have also developed assessments using games or problem-based complex scenarios in STEM problem solving, mainly emphasizing science education (e.g., ALCHEMIST, Use Your Brainz, ChemLabBuilder) (Annaggar & Tiemann, 2020; Scherer & Tiemann, 2012; Shute et al., 2016). These assessments raised issues of the lacking role and underrepresentation of mathematics in STEM assessments (Lasa et al., 2020; Maass et al., 2019).

Another problem, teachers encounter difficulties in implementing STEM assessment due to a complex area of assessment, the lack of applicable resource, and practical problems (Amalina & Vidákovich, 2022b; Gao et al., 2020). Assessing students in STEM area is challenging since it needs to connect several disciplines according to the curriculum and framework, but in fact, there is a lack of explicitly and well-articulated framework in this area (Amalina & Vidákovich, 2022b). In addition, there are still lack of assessment tools that are applicable in the class (Amalina & Vidákovich, 2022b; Gao et al., 2020). They mostly research-oriented and time-consuming assessments. The constructed type test is the most applicable test for measuring cognitive in STEM education, but the psychometric evidence of the available tests are mostly difficult to obtain (Amalina & Vidákovich, 2022b).

Although integrating STEM in problem solving has been proven essential in middle school mathematics, the trend of STEM assessment is still monodisciplinary, specifically in the area of mathematical problem solving assessment in Indonesia (Amalina & Vidákovich, 2022b; Gao et al., 2020). However, the Indonesian 2013 mathematics curriculum requires students to master mathematical problem solving related to daily life and STEM integration. Consequently, students encounter difficulties in solving interdisciplinary problems (Suratno et al., 2020).

According to several issues raised by researchers, a test that emphasizes mathematical problem solving based on an integrated STEM framework for Indonesian middle school students should be developed. The test focuses on interdisciplinary knowledge and skills (cognitive) as a part of STEM activity assessment through problem solving. A mathematical task is situated in the written science context that requires students to apply their knowledge about mathematics, science, and engineering-based design using a specific tool to solve the problem. Therefore, in this study, we develop an integrated STEM-based mathematical problem-solving test, validate the content and construct of the developed test, and examine the reliability of the developed test using inter-rater reliability.

Framework in STEM Problem Solving

The definition of STEM education is an interdisciplinary learning approach that involves knowledge integration or combining two or more disciplines among science, technology, engineering and mathematics (Gao et al., 2020; Jolly, 2016). Science entails applying content, context, or a manner of thinking that provides a context for reflection, organization, and action. Mathematics provides a set of notions, skills, and contents for solving problems (Jolly, 2016; Lasa et al., 2020). Engineering offers an engineering-based design that enables the model construction (Lasa et al., 2020). Additionally, technology refers to technical tools that ease a process (Jolly, 2016; Lasa et al., 2020).

Infusing STEM approach in the education area is required since it includes the real-world context and problem solving (Maass et al., 2019; Tasir et al., 2018). Solving an interdisciplinary problem is a key component of STEM education that requires problem solving skills (Jolly, 2016). Problem solving provides a powerful context for learning and practicing STEM concepts, and draws on skills, approaches, and ways of thinking that can be applied across disciplines (Priemer et al., 2020). The process of solving STEM problem depends on the type of problem (Jolly, 2016). A problem that requires students to produce an artefact to demonstrate their mastery of content called problem-based activity (Jolly, 2016; Priemer et al., 2020). However, there is a type of STEM problem that requires students to apply their knowledge based on authentic STEM text or scenario (Shanta, 2019). An integrated STEM based-mathematical problem is not a problem-based activity. It is a mathematical problem in a science scenario that developed by integrating problem-solving framework in every discipline.

The available STEM problem solving frameworks are based on frameworks in every STEM discipline. In science, problem-solving is always embedded in the scientific inquiry framework and scientific reasoning (Priemer et al., 2020). Problem-solving in science emphasizes on inductive thinking, while in mathematics allows inductive and deductive thinking. Problem-solving in mathematics is a process to solve mathematical tasks that potential to provide intellectual challenge by understanding a problem and applying mathematical concepts to generate a conclusion (Lasa et al., 2020; Priemer et al., 2020). The preliminary mathematical problem-solving framework is heuristics that continuously followed by other framework related to it (Priemer et al., 2020). Researchers have exploited mathematics approaches closely related to scientific inquiry, such as mathematics experimentation and inquiry-based mathematics education. The opposite of the concept of "inquiry" is "proving" (Priemer et al., 2020). The problem solving framework in engineering is related to engineering design or creating a solution (Jolly, 2016; Kelley & Knowles, 2016; Wells, 2016).

The role of mathematics in STEM problem-solving is to provide a set of notions and skills that allow solving problems, including mathematical ability to analyze, reason, and interpret solutions (Jolly, 2016; Lasa et al., 2020). Another skill that has an important role in STEM problem-solving is mathematical modelling since STEM problem is a complex problem that requires to model from a real-world problem into a mathematical sentence (Maass et al., 2019). In addition, mathematics engages in STEM problem-solving by application of mathematics concepts and procedures to the problem (Lasa et al., 2020).

Several researchers also have developed STEM frameworks by applying frameworks in various disciplines, focusing on practical problem solving (Kelley & Knowles, 2016; Wells, 2016). Priemer et al. (2020) developed framework using a PISA framework as a ground framework and embedded some domain-specific problem-solving techniques, e.g., scientific inquiry, proving, engineering design, and computer science. Amalina and Vidákovich (2022a) developed a framework by modifying Priemer et al.'s (2020) framework to integrate STEM into mathematical problem-solving for summative and diagnostic assessments (see Table 1). We refer to Amalina and Vidákovich (2022a) framework to develop a test.

Table 1. An integrated STEM problem-solving framework

	Indicators
Exploring and understanding	(1) Identifying relevant and valuable unknowns and given information (STEM); (2) Determining the goal of a problem (STEM); (3) Employing a useful basic concept (SM)
Representing and formulating	(4) Constructing a tabular, graphical, symbolic, or verbal representation of a problem using technology (STM); (5) Making a hypothesis, developing criteria, or checking existing theory (STEM).
Planning and executing	(6) Formulating a reasonable argument, explanation, and solution (including strategy, step design, and model building) (TEM); (7) Arranging, critically choosing, and evaluating alternatives (STEM); (8) Performing a plan by deducting, proofing, or finding a counterexample, applying mathematics concepts, mathematization, reasoning, computational skills, science concept, and technology (STEM).
Monitoring and reflecting	(9) Drawing conclusions, evaluating, and reflecting on results and methods (STEM).

Note: S = science, T = technology, E = engineering, M = mathematics

Content and Context in STEM Education

Based on the analyses of the Cambridge curriculum, Indonesian curriculum, PISA framework, and literature review, mathematics contents used in STEM tasks are mostly measurements, geometry, and arithmetic. Moreover, the contexts used are mainly scientific contexts related to the environment and how to protect it. These contents and contexts can be integrated into STEM problem solving and real-world problems (Lasa et al., 2020). There are several content knowledge intersections in every grade.

1. Mathematics: Arithmetic (number, fraction, and measurement), Algebra (ratio and proportion), Geometry (two-dimensional (2D) figure), and Statistics (data representation and central tendency)
2. Science: Pollution, climate and disaster, eco-friendly product, electricity, and energy.

There is no specific STEM curriculum in Indonesia. Science and mathematics are taught separately at middle school (grades 7 to 9) or International Standard Classification of Education (ISCED) level 2. Technology serves as a tool in the teaching and learning process. Nevertheless, the importance of integrating STEM in the Indonesian mathematics classroom has been emphasized.

Although there are no differences in grades' main competencies, basic competencies may be different and higher in some grades. Some basic competencies in a certain grade can become the prerequisites for basic competencies in a higher grade. Problem-solving competencies are the compulsory competencies for applying mathematics concepts. Solving problems related to numbers and fractions are compulsory basic competencies in every grade. In Grade 7, students have to master problem solving in ratios, proportions, 2D figures, and data representations. Students in Grade 8 have to acquire knowledge of solving 2D and three-dimensional (3D) figures and central tendency problems. Grade 9 students should master the prior knowledge of function (represent data), 3D figure (including prior knowledge about circle), and congruency and similarity (ratio and proportion).

METHOD

Instrumentation

The instrument we used is an integrated STEM-based mathematical problem-solving test, a developed complex scenario essay test. There are three scientific contexts in every grade. Every scientific context has eight items (prompting items) to explore students' problem-solving skills using indicators in Table 1. The scoring used in every prompting item is from 0 (blank answer) to 5 (complete and correct answer). Students are allowed to use any tool or application to help them. The time given in every scenario is 1 h.

The tool for assessing content validity is an assessment sheet that experts will complete. It assesses scenarios, items, and scoring. It contains four aspects: sufficiency, clarity, coherence, and relevance aspects. The sufficiency aspect includes the suitability of aims, curriculum, indicators, and student level. The clarity aspect measures the language (syntax and semantics). The coherence aspect refers to the relatedness to dimension and indicator. The relevance aspect measures the importance level of an item. Experts must rate using four scales (from 1 = not clear to 4 = very clear). The qualitative assessment is also provided by giving suggestions on each item for improvement.

Procedure

The procedures for developing the test are: (1) analyze the research needs and develop a framework, (2) analyze the contents and contexts that could be applied in the test, (3) develop scenarios of the test related curriculum contents and contexts, (4) develop prompting items in every scenario based on the problem-solving indicators and basic competencies, and (5) generate the answer and scoring method.

After developing the instrument, content validation was performed by expert panels. The experts rated the instrument using the assessment sheet. The experts were recruited with criteria: (1) a minimum of a master's degree in mathematics education, (2) must be an educator, and (3) a minimum of three years of working experience. The qualitative assessment from experts in each item will be used to improve items. After revising based on the experts' suggestions, the test was administered to middle school students to analyze the construct validity. The data collection was performed between March 22 and April 12, 2022. The test was administered in 3 h using Google Forms.

Because the test is an essay, we need to ensure consistency in the implementation of a rating system. As such, two raters—one of the authors and a female teacher who has a master's degree in mathematics education—checked students' answers. The students' scores from two raters were compared and analyzed using statistical tests.

Participants

Three female experts were selected to assess content validity. They are secondary school mathematics teachers in an urban area who have a master's degree in mathematics education. The test was conducted on Grades 7–9 students in East Java, Indonesia, to check the psychometric evidence. We selected A-accreditation schools with random classes. Five schools participated, with a total of 286 participants from three areas. Table 2 summarizes the characteristics of the participants.

Table 2. Characteristics of participants

Demographic	Characteristics	N	%
Gender	Boys	102	35.66
	Girls	184	64.34
Grade	7 (M age = 13.09, SD = 0.61)	116	40.56
	8 (M age = 13.99, SD = 0.49)	90	31.47
	9 (M age = 14.95, SD = 0.48)	80	27.97

School location	City	128	44.76
	District	158	55.24
Ethnic	Javanese	279	97.55
	Madurese	4	1.40
	Others	3	1.05

Note: N = the number of students, SD = Standard deviation, M age = mean age

Data Analysis

Content validity was analyzed using the content validity index (CVI) and intraclass correlation coefficient (ICC). The CVI is a quantitative evaluation to measure the mean content validity ratio given by experts. The CVI is the most widely used index for content validation. The ICC is the correlation test to measure agreement among non-binomial scores given by experts. Since in the current study there are three experts rated the test by 1-4, the ICC is applicable to measure how similar are the scores given by experts.

The construct validity was examined using Rasch analysis, which described (1) the item fit model, (2) reliability, (3) item discrimination, (4) the item level of difficulties, (4) the Wright map, and (6) scale step analysis. The consistency of implementation of the rating scale (inter-rater reliability) was examined using ICC. The analysis was performed using SPSS 25 and Conquest applications.

The model used in the Conquest application is the partial credit model. The fit of this model was proven significantly better than the fit of the rating scale model according to the Conquest manual book. To explore the item fit model, we used the weighted (infit) mean square (MNSQ) and unweighted (outfit) MNSQ values. The ideal value of outfit and infit MNSQs is 1 based on the Rasch measurement model, but items are categorized as fit if the outfit or infit MNSQ values of [0.5, 1.6) (Andrich, 2018; Bond & Fox, 2015). The acceptable item discrimination value is more than 0.2.

A separation reliability coefficient evaluates whether the localization parameters of the items are sufficiently separate to cover the entire ability interval. The parameters are separate enough if the reliability coefficient is greater than or equal to 0.90 or $\chi 2$ is significant (López-Pina et al., 2016). The expected a posteriori/plausible value (EAP/PV) reliability is an estimate for test reliability provided by the Conquest software.

The ability parameters of the students are compared with the item localization parameters through a Wright map, which also shows the item-level difficulty. Moreover, the item-level difficulty is reported through estimate values. The higher the estimated values, the more difficult the item is. The scale step analysis aims to analyze the scoring function or behavior. The scale behavior is represented by the "Pt bis" value—the additional result of a category and correlation of those items— and the percentage of students who achieve every scale (0–5).

We applied Rasch analysis for both prompting items and scenarios. Scenarios act as items that construct the test, and a prompting item refers to a subitem of the test. Hence, it is necessary to examine the extent to which the test accurately assesses problem-solving skills through the scenarios and prompting items. We used the total score of students answering the prompting items in every scenario to analyze the fit model in every scenario. The total score is 40, but we recode it for the Conquest application. We recode them into a 1–5 scoring scale with the scoring rubric. We performed a Rasch analysis using the students' scores in every prompting item to analyze the fit model of prompting items. There are eight prompting items in every scenario. The total number of scenarios is six. Thus, there are 48 prompting items in total. The score used is from 1 to 5 in each prompting item.

RESULTS AND DISCUSSION

Developing an Integrated STEM-Based Mathematical Problem-Solving Test

The assessment trend in STEM is an essay scenario problem test that we used as a basis to develop a new test. The integrated STEM-based mathematical problem-solving test is a problem-based complex scenario test that integrates mathematics through the new STEM framework (Table 1). It is a mathematical scenario word problem related to scientific phenomena and applies engineering design to solve a problem with the help of technology. Mathematics is taking parts as contents, notions, and skills in problem solving. Science participates in the manner of thinking and the context used. Engineering engages in engineering-based design, and technology provides a tool used during problem solving.

There are three scenarios in every grade related to environmental management with one or two anchor scenarios in every grade. Every scenario has a challenge. The contents used are number and measurement, ratio and proportion, geometry, and statistics. Each scenario has eight dependent prompting items. The prompting items are developed from the sub indicators, integration between the indicators (Table 1) and basic competencies of middle school Indonesian 2013 curriculum. The prompting items in every topic are metacognitive prompting questions for guiding students to explore their problem-solving steps and reveal their thinking. The use of technology as aid is required (in this case, calculators, grid papers, rulers, and any possible devices are allowed).

The maximum score in each indicator is 5, adapted from Docktor et al. (2016) and Salazar-Torres et al. (2021). Score 5 for a complete and correct answer, 4 for a complete answer with minor error, 3 for incomplete but correct or complete with major error, 2 for incomplete answer with minor error, 1 for completely wrong and irrelevant or incomplete with major error, and 0 for a blank answer. The total maximum score for every scenario is 40. Table 3 explains the description of each scenario.

Table 3. Description of every scenario in the integrated STEM-based mathematical problem-solving test

Scenario	Grade	Contents	Descriptions
Eco-Friendly Packaging	7	Area (square and rectangle), simple arithmetic calculation, and decimal	Design eco-friendly packaging with simple constraints and decide on the lowest price
School Park	7	Ratio, area and circumference of rectangle, diameter, fraction, designing graph, and simple arithmetic calculation	Design a planted school park with the trees that can absorb the largest amount of CO_2 and represent the CO_2 absorbed in several years by graph
Calory versus Greenhouse Gas Emission	7, 8, and 9	Percentage, decimal, and simple arithmetic calculation	Design a 1-day-menu that fulfills the nutrition and has the lowest CO_2 emission during their productions
Flood Water Reservoir	8	Central tendency, table, simple arithmetic calculation, and volume	Design a flood water reservoir and decide the fastest time by pumps to absorb the average volume
City Park	8 and 9	Percentage, ratio, decimal, designing graph, diameter, complex arithmetic calculation, conversion, and rectangle area	Design a planted city park with the trees that can absorb the largest amount of CO_2 and represent the remaining CO_2 emission in several years by graph after planting
Infiltration Well	9	Diameter, ratio, circle, and complex arithmetic calculation	Design an infiltration well with complex constraints and decide the price to build it

The contents tested in "Calory versus Greenhouse Gas Emission" are the contents that the students already studied in Grade 7 and are basic competencies (as well as basic competencies of prior knowledge) in all grades. Therefore, the scenario could be tested for all grades. The contents in "City park" were already studied in Grades 7–9; however, because of complex constraints, they will be administered in Grades 8 and 9. The "Infiltration Well" contents are only suitable for Grade 9 because the cylinder (and the prior knowledge: circle) topic appears in the Grade 9 curriculum. Moreover, the "Flood Water Reservoir" could not be administered in Grade 7 because the central tendency topic appears in Grade 8. Figure 1 illustrates an example of a scenario in the test and the prompting items.

Scenario 1: Eco-Friendly Packaging

Eco-friendly and low-budget products are starting to be prioritized. We want to produce packaging with the following criteria.

1. The packaging has length, width, and height of 18 cm, 18 cm, and 8 cm, respectively
2. The cover of the packaging is square and 1.2 times the area of the base.
3. There are four material options:

	Duplex paper	Ivory paper	Styrofoam	Plastic
Size	79 cm × 109 cm	215 mm × 330 mm	100 cm × 50 cm	17 cm × 50 cm
Thickness	250 gsm	250 gsm	0.5 cm	0.5 cm
Price	Rp 4,000.00	Rp 2,000.00	Rp 9,000.00	Rp 1,300.00
Time for decomposing	2–6 months		Cannot decompose	1,000 years

Challenge

Design an eco-friendly packaging and decide on the lowest budget for packaging a product based on the criteria.

Prompting questions:

1. What is your challenge (asked) in the scenario?
2. What information do you need to resolve the challenge?
3. What is the total area of four sides, cover, and base in cm^2?
4. What is your guess relating to the type of material that could be used to design the packaging? Give your reasons.
5. Draw your packaging design based on the shape and length.
6. Mention two materials that are unsuitable to be chosen and give your reasons!
7. How much is the price for your packaging based on your chosen material and area of your packaging?
8. Is the packaging eco-friendly and the cheapest one? Give your conclusion about the price and type of material used to produce packaging.

Figure 1. Example of a scenario and prompting items in the developed test

Content Validity of the Developed Test

The CVI value for every scenario is 1. However, all experts agreed with a score of 4 (very clear) for only the "Eco-Friendly Packaging." In the "Calory versus CO2 Gas Emission" and "Infiltration Well" scenarios, all experts agreed with a score of 3 (clear) due to the need for minor revision regarding language clarity and sufficiency at the student level. The "City Park" and "School Park" scenarios received one 4 score and two 3 scores due to several items' problems. Moreover, in the "Flood Water Reservoir" scenario, due to language barrier, only an expert scored 3. See Table 4 for detailed numerical results.

Table 4. The content validity score for every scenario

Scenario	Expert 1's score	Expert 2's score	Expert 3's score
Eco-friendly packaging	4	4	4
School park	4	3	3
Calory versus gas emission	3	3	3
Flood water reservoir	4	4	3
City park	4	3	3
Infiltration well	3	3	3

The CVI values of the prompting items in every scenario ranged between 0.67 and 1 (see Table 5). An item with a CVI of 0.67 is an item that measured students' skills in generating a conclusion in every scenario. However, only one judge gave a score of 2, and two judges granted a score of 3. Judges stated that the item is somewhat relevant, but another item may be covering what this item is measuring.

Table 5. The content validity score for every prompting items

Expert 1	Expert 2	Expert 3	CVI	Prompting items
4	4	4	1	1, 3—7, 9, 11—15, 17, 19—23, 25, 27—31, 33, 35—39
3	3	4	1	10, 41—47
3	4	4	1	2, 18
3	2	3	0.67	8, 16, 24, 32, 40, 48
4	3	4	1	26, 34

The experts agreed that the scenarios and prompting items suffice to measure student level based on curriculum and indicators and were related to the measuring dimension or indicator. However, several terms and sentences needed minor revision. The Cronbach alpha indicated acceptable with .835, and the ICC result represented moderate reliability (rxx = 0.628).

Qualitative recommendations are suggested by judges: (1) reduce the item and the hints in the prompting items, (2) change the scenarios' format, and (3) make a uniform score. The previous part's description of the test development is the revision version after the experts' assessment.

Construct Validity of the Developed Test

General Analysis Based on Scenario

The results of Rasch analysis showed 52 iterations, 27 estimated parameters, and 1,800.663 final deviance. The weighted MNSQ value ranged between 0.84 and 1.12, with various scenario difficulty levels. Scenario 2 was the most difficult (estimate value = 1.307) and scenario 4 was the easiest (estimate value = −1.141). Table 6 describes the response model parameter estimates in every scenario.

Table 6. Response model parameter estimates based on scenario

Scenario	Estimate	Error	Unweighted MNSQ	Weighted MNSQ
1. Eco-Friendly Packaging	0.610	0.138	0.86	0.89
2. School Park	1.307	0.135	0.86	0.84
3. Calory versus Gas Emission	−0.156	0.106	1.22	1.12
4. Flood Water Reservoir	−1.141	1.148	1.10	0.93
5. City Park	−0.986	0.123	0.98	0.94
6. Infiltration Well	0.366	0.293	0.96	0.96

Rasch analysis also measured the difficulty level for students to make an achievement transition from one step to another. In this case, the score used in every scenario is from 0 to 5. Hence, the Rasch analysis calculated the transition among these scores. The results that revealed the hardest transition in all scenarios was from score 4 to 5 and the easiest transition was from score 1 to 2 or 0 to 1, indicating that students found it easier to reach score 2 when they had score 1 and students found it more difficult to reach score 5 when they had score 4. Scenario 3 had the most difficult transition from score 4 to 5 compared with other scenarios, with an 8.968 estimate value. Additionally, scenario 3 had the easiest transition from score 0 to 1 compared with the others, with a −7.572 estimate value. Figure 2 (a) depicts the Wright map based on the scenario.

The discrimination values of the scenarios ranged between 0.85 and 0.94. Pt bis was constantly increasing, and toward the last values, it was positive. For instance, the Pt bis values of scores 0–5 in scenario 5 are −0.49, −0.37, −0.37, −0.31, 0.12, and 0.59, respectively. The percentage of students who achieved every scale clarified the equal distribution, indicating that the behavior of the scoring scale is acceptable and fit. The separation reliability was high at 0.984 (χ^2 (5)=239.341,p<.001), and the EAP/PV reliability was 0.867. The high separation reliability indicates that all scenarios can cover all students' abilities. Furthermore, the high EAP/PV reliability shows that the test is reliable.

Detailed Analysis Based on Prompting Items

Rasch analysis performed 158 iterations, 224 total estimated parameters, and a final deviation of 14,130.920. The weighted MNSQ values ranged between 0.66 and 1.53. Four prompting items had unweighted MNSQ values of more than 1.60, namely, items 4, 19, 36, and 38. Nevertheless, the weighted MNSQ values of all prompting items indicated fit. Item 4, regarding making hypotheses in the "Eco-Friendly Packaging" scenario, had an unweighted MNSQ value of 2.63. Item 19, about using the basic concept in the "Calory versus Gas Emission" scenario, had a 1.82 unweighted MNSQ value. Items 36 and 38 request students to make a hypothesis and arrange critical choices in the "City Park" scenario, which had 1.62 and 1.92 unweighted MNSQ values, respectively.

The difficulty levels were various, with item 25 as the easiest prompting item (estimate value = −1.766) and item 47 as the most difficult prompting item (estimate value = 2.065). Item 25 requires students to determine the goal in the "Flood Water Reservoir" scenario. Meanwhile, item 47 requests students to apply the concept in the "Infiltration Well" scenario.

Regarding the difficulty levels for achievement transition from one score to another, the results revealed that the most difficult levels for students are to achieve from score 3 to 4 and score 4 to 5. Seven prompting items (Items 4–6, 11, 12, 36, and 46) reported that the most difficult transition of student achievement was from score 3 to 4. The 41 other prompting items reported that the most difficult transition of student achievement was from score 4 to 5. The hardest transition was prompting Item 21 from score 4 to 5 (estimate value = 6.333).

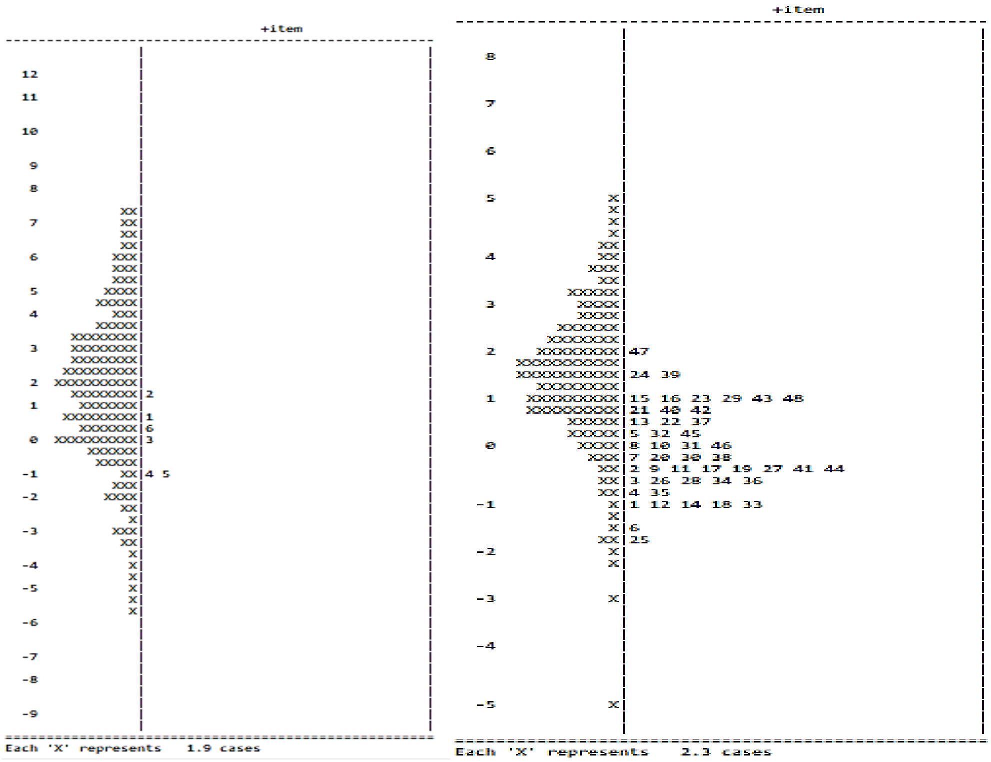
The easiest achievements were from score 0 to 1, score 1 to 2, and score 2 to 3. Four items reported that the easiest transition was from score 2 to 3, namely, items 41–43 and 47. Twelve prompting items (items 1, 2, 9, 10, 17, 26, 34, 40, 44–46, and 48) notified that the easiest transition was from score 1 to 2. The other prompting items showed that the student achievement transition from 0 to 1 was the easiest. The easiest transition was reported on item 21 from score 0 to 1 (estimate value = −6.539). Figure 2 (b) explains the Wright map of the distribution of student achievement based on items.

The discrimination values of the items ranged between 0.64 and 0.87. Pt bis was constantly increasing, and toward the last values, it was positive. The percentage of students who achieved every scale clarified equal distribution. The separation reliability was high at 0.986 (χ^2 (47)=3,411.843,p<.001) and the EAP/PV reliability was 0.962. The high separation reliability indicates that all prompting items can cover all students' abilities. Furthermore, the high EAP/PV reliability shows that the test is reliable.

Furthermore, we discuss the misfit items. The items are misfit because the unweighted MNSQ value is more than the expected value. The unweighted MNSQ value is less important since Conquest analyzed the observed curve (curve from empirical data) based on the weighted MNSQ results compared with the model curve that has an MNSQ value of 1. According to the results, the observed curves of these misfit items were a little flatter than the model curve (underfit). This will often be the case when the MNSQ is significantly greater than 1, indicating that the data were less predictable than the model expected or the data had more variation in the observed pattern response. For instance, the weighted MNSQ for item

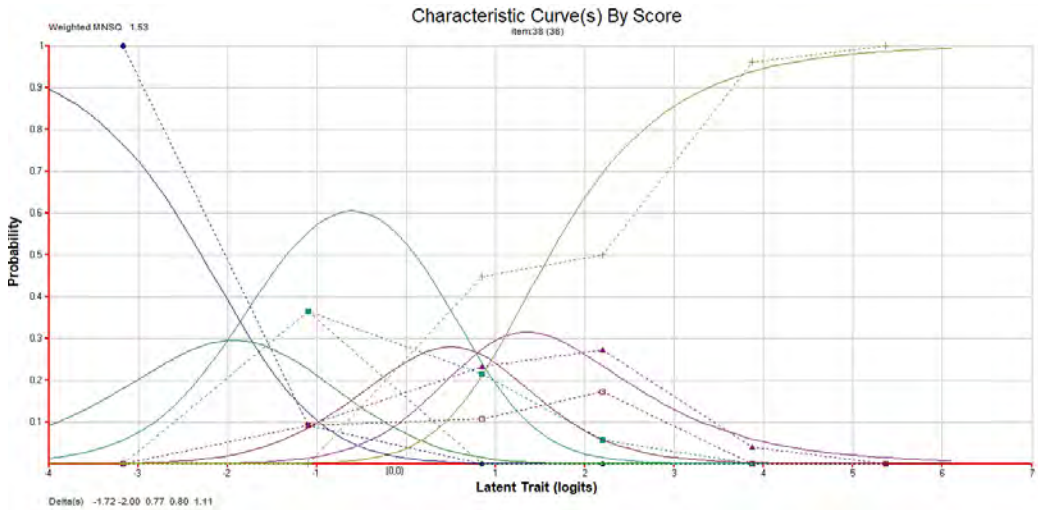
38 was 1.53, indicating that there was 53% more randomness (i.e., noise) in the data than modeled. Hence, further investigation is required to determine the reason.

Figure 3 shows the characteristic curve for each category (five scores) for item 38. Empirical and modeled curves are matched via color. The disparity between the observed and modeled curve for score 2 and 3 were the largest, and this was consistent with the high fit statistics for this category (weighted MNSQ value more than 1.6), meaning that students who received a score of 2 and 3 behaved differently when answering Item 38.



Note for scenario: Item 1: Eco-Friendly Packaging; Item 2: School Park; Item 3: Calory versus Gas Emission; Item 4: Flood Water Reservoir; Item 5: City Park; Item 6: Infiltration Well.

Figure 2. (a) Item (scenario) and students' ability distribution map. (b) Item (prompting item) students' ability distribution map.



Note: Scores 1–5: blue, green, light blue, red, and light brown, respectively.

Figure 3. Item characteristic curve for Item 38

Similar patterns were found for items 4, 19, and 36. Students who received a score of 2 and 3 behaved differently. It was observed from the curve and confirmed from the highest fit statistics compared with other categories. The weighted MNSQ values for score 2 and 3 in items 4, 19, and 36 were above 1.6.

Reliability of the Integrated STEM-Based Mathematical Problem-Solving Test

Students' answers (n = 286) to the test were checked by two experts. The mean score of every student from the two experts was analyzed using ICC. The results revealed a good ICC (rxx = 0.958) and good reliability (α = 0.979). According to the result, the consistency of the implementation of a rating scale was proven.

Discussion

Assessment in interdisciplinary problem solving, specifically in STEM area, is required to tackle 21st-century challenges. However, issues related to underrepresented mathematics, the trend of monodisciplinary problem solving in Indonesia, and the lack of diagnostic tests in this area have been detected (Lasa et al., 2020; Maass et al., 2019). Hence, the development of an integrated STEM-based mathematical problem-solving test resolves these issues.

The contents and contexts of the test were developed based on the Indonesian 2013 curriculum. An essay scenario-based problem was chosen as the type of test. The essay test is suitable to be applied to assess students' problem-solving, which is process-oriented rather than result-oriented (Gao et al., 2020). Scenario-based problem is the most common type of test in interdisciplinary problem solving because it embodies the authenticity of a problem that engages students in it (Gao et al., 2020; Shanta, 2019; Tasir et al., 2018). The anchor scenario was raised from the intersection of basic competencies that appear between grades to compare the problem-solving achievement of Grades 7–9.

Metacognitive skills are implied and practiced during mathematical problem solving. Metacognition in problem solving focuses on planning, monitoring, evaluating, and reflection (Velasquez & Bueno, 2019). Metacognitive skills cannot be assessed explicitly. Hence, metacognitive prompt questions are required. Metacognitive prompting items in the developed test are designed to reveal student thinking, ensure student responses were consistent, address the students' abilities identified, and explore every problem-solving indicator. It agrees with the study aim of Shanta (2019) that uses metacognitive prompts in his developed test.

We performed content validity, construct validity, and reliability assessments to ensure the psychometric appropriateness of the developed test. Practitioners were chosen to assess content validity because they understand the real condition of the students and curriculum. The result of ICC indicated moderate (Koo & Li, 2016). The CVI was adequate for all scenarios, but several prompting items had a CVI value of less than 0.69. CVI is categorized as adequate if it is greater than 0.69; however, if 50% of the judges gave a score of 3 or 4, a CVI value of less than 0.69 is still acceptable (Hyrkäs et al., 2003). In this case, the prompting items that had a CVI value of 0.67 were generated from two scores of 3 and one score of 2. Because more than 50% of judges (two of three) gave a score of 3, the CVI values of these prompting items were acceptable. The low CVI value was because of the few number of judges.

The result of construct validity revealed all scenarios were categorized as fit based on chosen theories (Andrich, 2018; Bond & Fox, 2015). The discrimination items, behavior of the score, and reliabilities were acceptable (Bond & Fox, 2015). The most difficult scenario was the "School Park." The most common difficulty was that students could not differentiate between area and circumference concepts that they want to use, which agrees with the previous research that there is misalignment between geometry item difficulty and students' abilities (Bostic & Sondergeld, 2015). Moreover, students failed to make hypotheses and decide the best option from several options. Providing several options perhaps increases student difficulty in solving the scenario. The easiest scenario was "Flood Water Reservoir" because of the simple constraint of the available options.

The achievement transition from score 4 to 5 was the most difficult because score 5 requires a perfect answer. The easiest transition is from score 0 to 1 because score 0 is for a blank answer and score 1 is for an irrelevant answer or wrong answer. It was rare for students in all phases of problem solving to receive a score of 0. Hence, the probability of receiving more than a score of 0 is high.

A detailed analysis was performed to check the model fit of prompting items. According to the weighted MNSQ results, all prompting items were fit (Andrich, 2018; Bond & Fox, 2015). Conquest analyzed the observed curve based on the weighted MNSQ results compared with the model curve. Hence, fit items can be detected only with weighted MNSQ, but it will be more accurate to observe both weighted and unweighted MNSQs. However, the unweighted MNSQ results for four prompting items were misfit. The unweighted MNSQ is more sensitive to responses to items with difficulty far from a person or vice versa (Bond & Fox, 2015). These are usually easy to remedy and diagnose (e.g., lucky guess or careless mistake).

According to item characteristic curves, the data from students who received a score of 2 and 3 caused misfits in these items. We decided not to delete these items because it will distort the test's construct. Items 4, 36, and 38 requested students to make hypotheses and eliminate impossible answers. After checking the students' answers, mostly they were confused between "possible" and "impossible" words as well as "guess or hypotheses" words. Hence, they put "possible guess" into the "impossible guess." However, based on their general answers, they understood how to make hypotheses and evaluate alternatives. Therefore, it can be concluded that the misfit items were because of the language barrier. Item 19 is about using the basic concept used in solving a problem, in this case, calculating the required protein, fat, and carbohydrate. The common mistakes were (1) students only mentioned the required nutrition (in percentage) based on the given information without calculating them, and (2) some students interpreted the word "fulfill" in the question with "more than," but the rest interpreted it as "less than." However, after examining their complete problem-solving process, they understood how to calculate them. They misinterpreted the question. This is perhaps because of an ambiguous prompting question. The original prompting item was "How many kcal of fat, carbohydrate, and protein are needed to fulfill the criteria?" and it was revised to "calculate the fat, carbohydrate, and protein needed based on the criteria in kcal." Additionally, we modified the word "fulfill" into "more than."

The discrimination items, behavior of the score, and reliabilities in the prompting items were acceptable (Bond & Fox, 2015). The most difficult item was item 47, which requested the students to calculate the price to build an infiltration well. The most common error was that students could not differentiate between the concept of surface area and volume that they wanted to use, which agrees with the result of Tan Sisman and Aksu (2016). Additionally, item 47 requires complex skills and concepts, e.g., number and fraction, 3D figure, ratio and proportion, and circle. Item 25 is the easiest because it only needs students to understand the already stated question's goal. The difficulty levels for transiting achievement in the prompting items were similar to the difficulty levels for transiting achievement in the scenario, with the transition from score 4 to 5 as the most difficult and 0 to 1 as the easiest transition.

Inter-rater reliability is necessary for measuring the consistency of the implementation of a rating score in essay tests. The reliability was certified as good (Koo & Li, 2016).

CONCLUSION

The development of a mathematical problem-solving test based on an integrated STEM framework addressed the issues of previous studies with emphasis on mathematics while still ensuring interdisciplinarity. Six essay scenario-based problems with anchor scenarios in every grade were developed. Metacognitive prompting items were provided in every scenario based on indicators. The developed test was proven to be adequate based on experts' judgments and content assessment.

The construct validity assessment was performed using the Rasch model, which indicated that all scenarios were fit with a reasonable difficulty level. The behavior of the score, EAP/PV reliability, and discriminant value were acceptable. Additionally, the weighted MNSQ value, discriminant value, and EAP/PV reliability of all prompting items denoted fit and acceptable. However, four prompting items showed misfit according to unweighted MNSQ values.

The suggested revision of misfit items was emphasized in the language aspect. The suggested revision is formulated according to the result of the item characteristic curve and the most common errors encountered by students.

This study has limitations regarding the test and the sampling method. The developed test was only administered using Google Forms because of the lack of students' knowledge of other platforms. Additionally, the roles of engineering and technology in the test are underrepresented compared to mathematics and science. Regarding the sample, it is suitable to involve students from different schools (e.g., sample Grade 8 was only from a school). Only involving students from a school limits the external validity.

The results of the current study will be beneficial for mathematics teachers in assessing students mathematical problem-solving skills in an interdisciplinary context. Additionally, the test can be used as well for science teachers for assessing specific science topics. Teachers can use the test as an assessment in interdisciplinary knowledge and skills as a part of STEM activity assessment. This test can be used as a preliminary test before conducting an action or experiment in a mini laboratory to prompt students' knowledge of an interdisciplinary STEM topic. Hence, before practicing in an activity, students have a plan regarding what they want to do and interdisciplinary understanding of the issue related to environmental management. The test emphasized on process rather than the product itself. Teachers can extend the test into practice to assess a product based on what students planned in the test.

For researchers in the relevant area, the results of the test validity can be a basis for adapting the test in a different sample background and for further testing of its validity. The future study is widely open as well regarding providing the test in a computer-based format.

Declarations

Author Contribution	:	IKA: Conceptualization, Writing, Formal analysis, Methodology, Editing and Visualization. TV: Review, Validation and Supervision
Funding Statement	:	This article received a funding from the Doctoral School of Education, University of Szeged, Hungary.
Conflict of Interest	:	No potential conflict of interest was reported by the authors.
Additional information	:	Additional information regarding the test is available from the first author

REFERENCES

Amalina, I. K., & Vidákovich, T. (2022a). A mathematical problem-solving framework-based Integrated

STEM: Theory and practice. *International Journal of Trends in Mathematics Education Research*, *5*(1), 1–11. https://doi.org/10.33122/ijtmer.v5i1.105

Amalina, I. K., & Vidákovich, T. (2022b). Assessment in STEM Problem-Solving: A Systematic Review. *The International Journal of Assessment and Evaluation*, *29*(2), 64–80. https://doi.org/https://doi.org/10.18848/2327-7920/CGP/v29i02/63-80

Andrich, D. (2018). Advances in social measurement: A Rasch measurement theory. In F. Guillemin, A. Leplège, A. Briançon, E. Spitz, & J. Coste (Eds.), *Perceived health and adaptation in chronic disease* (pp. 66–91). https://psycnet.apa.org/doi/10.1201/9781315155074-7

Annaggar, A., & Tiemann, R. (2020). A video game as an assessment tool of problem-solving competence. *Gesellschaft Fur Didactik Der Chemie Und Physik (GDCP)*, *April*, 377–380.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model : Fundamental Measurement in the Human Sciences Third Edition*. Routledge.

Bostic, J. D., & Sondergeld, T. A. (2015). Measuring Sixth-Grade Students' Problem Solving: Validating an Instrument Addressing the Mathematics Common Core. *School Science and Mathematics*, *115*(6), 281–291. https://doi.org/10.1111/ssm.12130

Docktor, J. L., Dornfeld, J., Frodermann, E., Heller, K., Hsu, L., Jackson, K. A., Mason, A., Ryan, Q. X., & Yang, J. (2016). Assessing student written problem solutions: A problem-solving rubric with application to introductory physics. *Physical Review Physics Education Research*, *12*(1), 1–18. https://doi.org/10.1103/PhysRevPhysEducRes.12.010130

Du Toit, S., & Du Toit, G. (2013). Learner metacognition and mathematics achievement during problem-solving in a mathematics classroom. *The Journal for Transdisciplinary Research in Southern Africa*, *9*(3), 505–518. https://doi.org/10.4102/td.v9i3.194

Gao, X., Li, P., Shen, J., & Sun, H. (2020). Reviewing assessment of student learning in interdisciplinary STEM education. *International Journal of STEM Education*, *7*(1). https://doi.org/10.1186/s40594-020-00225-4

Hyrkäs, K., Appelqvist-Schmidlechner, K., & Oksa, L. (2003). Validating an instrument for clinical supervision using an expert panel. *International Journal of Nursing Studies*, *40*(6), 619–625. https://doi.org/10.1016/S0020-7489(03)00036-1

İncebacak, B. B., & Ersoy, E. (2016). Problem Solving Skills of Secondary School Students. *China-USA Business Review*, *15*(6), 275–285. https://doi.org/10.17265/1537-1514/2016.06.002

Jolly, A. (2016). *STEM by design: Strategies and activities for grades 4-8*. Routledge. https://doi.org/10.4324/9781315679976

Jones, I., Swan, M., & Pollitt, A. (2015). Assessing Mathematical Problem Solving Using Comparative Judgement. *International Journal of Science and Mathematics Education*, *13*(1), 151–177. https://doi.org/10.1007/s10763-013-9497-6

Junpeng, P., Marwiang, M., Chiajunthuk, S., Suwannatrai, P., Chanayota, K., Pongboriboon, K., Tang, K. N., & Wilson, M. (2020). Validation of a digital tool for diagnosing mathematical proficiency. *International Journal of Evaluation and Research in Education*, *9*(3), 665–674. https://doi.org/10.11591/ijere.v9i3.20503

Karatas, I., & Baki, A. (2013). The Effect of Learning Environments Based on Problem Solving on Students ' Achievements of Problem Solving. *International Electronic Journal of Elementary Education*, *5*(3), 249–267. https://www.iejee.com/index.php/IEJEE/article/view/25

Ke, F., & M. Clark, K. (2020). Game-Based Multimodal Representations and Mathematical Problem Solving. *International Journal of Science and Mathematics Education*, *18*(1), 103–122. https://doi.org/10.1007/s10763-018-9938-3

Kelley, T. R., & Knowles, J. G. (2016). A conceptual framework for integrated STEM education. *International Journal of STEM Education*, *3*(1). https://doi.org/10.1186/s40594-016-0046-z

Kesorn, N., Junpeng, P., Marwiang, M., Pongboriboon, K., Tang, K. N., Bathia, S., & Wilson, M. (2020). Development of an assessment tool for mathematical reading, analytical thinking and mathematical writing. *International Journal of Evaluation and Research in Education*, *9*(4), 955–962. https://doi.org/10.11591/ijere.v9i4.20505

Khairani, A. Z., & Sahari Nordin, M. (2011). The Development and Construct Validation of the Mathematics Proficiency Test for 14-Year-Old Students. *Asia Pacific Journal of Educators and Education*, *26*(1), 33–50. http://web.usm.my/apjee/webtest/APJEE_26.1.2011/26.1.2011_33-50.pdf

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lasa, A., Abaurrea, J., & Iribas, H. (2020). MATHEMATICAL CONTENT on STEM ACTIVITIES. *Journal on Mathematics Education*, *11*(3), 333–346. https://doi.org/10.22342/JME.11.3.11327.333-346

López-Pina, J. A., Meseguer-Henarejos, A. B., Gascón-Cánovas, J. J., Navarro-Villalba, D. J., Sinclair, V. G., & Wallston, K. A. (2016). Measurement properties of the brief resilient coping scale in patients with systemic lupus erythematosus using rasch analysis. *Health and Quality of Life Outcomes*, *14*(1), 1–8. https://doi.org/10.1186/s12955-016-0534-3

Maass, K., Geiger, V., Ariza, M. R., & Goos, M. (2019). The Role of Mathematics in interdisciplinary STEM education. *ZDM - Mathematics Education*, *51*(6), 869–884. https://doi.org/10.1007/s11858-019-01100-5

Priemer, B., Eilerts, K., Filler, A., Pinkwart, N., Rösken-Winter, B., Tiemann, R., & Zu Belzen, A. U. (2020). A framework to foster problem-solving in STEM and computing education. *Research in Science and Technological Education*, *38*(1), 105–130. https://doi.org/10.1080/02635143.2019.1600490

Salazar-Torres, J., Rincón Leal, O., & Vergel Ortega, M. (2021). The rubric as an assessment tool for solving problem situations in the physics and mathematics teaching context. *Journal of Physics: Conference Series*, *1981*(1). https://doi.org/10.1088/1742-6596/1981/1/012018

Scherer, R., & Tiemann, R. (2012). Factors of problem-solving competency in a virtual chemistry environment: The role of metacognitive knowledge about strategies. *Computers and Education*, *59*(4), 1199–1214. https://doi.org/10.1016/j.compedu.2012.05.020

Shanta, S. (2019). Developing and assessing authentic problem-solving skills in high school pre-engineering students. *ASEE Annual Conference and Exposition, Conference Proceedings*. https://doi.org/10.18260/1-2--32637

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior*, *63*, 106–117. https://doi.org/10.1016/j.chb.2016.05.047

Suratno, Wahono, B., Chang, C. Y., Retnowati, A., & Yushardi, Y. (2020). Exploring a Direct Relationship between Students' Problem-Solving Abilities and Academic Achievement: A STEM Education at a Coffee Plantation Area. *Journal of Turkish Science Education*, *17*(2), 211–224. https://doi.org/10.36681/tused.2020.22

Tan Sisman, G., & Aksu, M. (2016). A Study on Sixth Grade Students' Misconceptions and Errors in Spatial Measurement: Length, Area, and Volume. *International Journal of Science and Mathematics Education*, *14*(7), 1293–1319. https://doi.org/10.1007/s10763-015-9642-5

Tasir, Z., Mohamad, M. N. M., Ismail, Z., & Mustafa, N. (2018). Development and Validation of Problem Solving Task Based-Integrated STEM. *Proceedings - 2018 6th International Conference on Learning and Teaching in Computing and Engineering, LaTiCE 2018*, 17–23. https://doi.org/10.1109/LaTICE.2018.00011

Velasquez, Q. F., & Bueno, D. C. (2019). Metacognitive Skills in Problem Solving of Senior High School STEM Strand Students Metacognitive Skills in Problem-Solving among Senior High School STEM Strand Students. *Institutional Multidisciplinary Research and Development*, *2*(June), 124–129. https://doi.org/10.13140/RG.2.2.18820.99207

Wells, J. G. (2016). PIRPOSAL model of integrative STEM education: Conceptual and pedagogical framework for classroom implementation. *Technology and Engineering Teacher*, *75*(6), 12–19. http://america.aljazeera.com/articles/2013/9/16/many-us-bridges-arestructurallyunsoundsaysnewreport.html

Widodo, S. A., Ibrahim, I., Hidayat, W., Maarif, S., & Sulistyowati, F. (2021). Development of Mathematical Problem Solving Tests on Geometry for Junior High School Students. *Jurnal Elemen*, *7*(1), 221–231. https://doi.org/10.29408/jel.v7i1.2973