

Task-Level Variation in Elementary Students' Scientific Literacy Performance

Linda Hania Fasha^{1✉}, Wahyu Sopandi², Atep Sujana³, Muslim⁴

(^{1,2,3,4})Pendidikan Dasar, Universitas Pendidikan Indonesia)

DOI: <https://doi.org/10.28926/pej.v6i3.3284>

Abstrak

Literasi sains umumnya dilaporkan dalam bentuk skor keseluruhan atau skor setiap kompetensi. Namun, pelaporan tersebut dapat mengaburkan perbedaan kinerja siswa pada tugas yang mengukur kompetensi yang sama. Penelitian ini bertujuan menganalisis kinerja literasi sains siswa sekolah dasar pada tingkat butir soal serta mengidentifikasi variasi pada tiga kompetensi literasi sains. Penelitian menggunakan desain deskriptif-eksploratif dengan melibatkan 402 siswa dari tiga sekolah dasar. Peserta mengerjakan 10 butir soal dengan penskoran dikotomis yang mengukur kemampuan menjelaskan fenomena secara ilmiah, menginterpretasikan data dan bukti secara ilmiah, serta mengevaluasi dan merancang penyelidikan ilmiah. Data dianalisis pada tingkat keseluruhan, kompetensi, dan butir soal menggunakan distribusi skor dan proporsi jawaban benar. Hasil menunjukkan proporsi jawaban benar sebesar 64,0%, 74,7%, dan 72,6%. Pada tingkat butir, proporsi jawaban benar berkisar 32,4%–89,7%. Temuan menunjukkan bahwa asesmen perlu mempertimbangkan karakteristik dan tuntutan setiap tugas, bukan hanya kompetensi yang diukur secara keseluruhan.

Kata Kunci: literasi sains; variasi tingkat tugas; siswa sekolah dasar.

Abstract

Scientific literacy is commonly reported as an overall score or as scores for individual competencies. However, this form of reporting may obscure differences in students' performance across tasks designed to measure the same competency. This study aimed to analyze primary school students' scientific literacy performance at the item level and identify patterns of variation across and within three scientific literacy competencies. A descriptive-exploratory design was employed, involving 402 students from three primary schools. The participants completed 10 dichotomously scored items assessing their ability to explain phenomena scientifically, interpret data and evidence scientifically, and evaluate and design scientific inquiry. The data were analyzed at the overall, competency, and item levels using score distributions and proportions of correct responses. This analysis was followed by a conceptual examination of the demands of each task. The results showed that the proportions of correct responses across the three competencies were 64.0%, 74.7%, and 72.6%, respectively. Greater variation was found at the item level, where the proportion of correct responses ranged from 32.4% to 89.7%. Substantial variation was also observed among items assessing the same competency. For example, within the competency of explaining phenomena scientifically, the proportion of correct responses ranged from 32.4% to 89.7% across items. These findings indicate that competency-level scores may not fully capture variations in students' performance across individual tasks. Therefore, scientific literacy assessments should consider not only the competencies being measured but also the characteristics and demands of the tasks used to assess those competencies.

Key Words: scientific literacy; task-level variation; elementary school students

✉ Corresponding author : Linda Hania Fasha

Email Address : lindahaniafasha@upi.edu

Received 15 July 2026, Revised: 24 August 2026, Accepted: 17 September 2026, Published: 30 September 2026

Background

Assessment results should not only indicate how high or low students' ability is, but should also help explain which parts have been mastered and which remain difficult. In science literacy assessment, this information is often summarized as a total score or as achievement on each competency. Such summaries are convenient for describing overall performance, but they can obscure differences at the task level. Students with the same score do not necessarily succeed on the same type of item. Likewise, several items measuring similar competencies can produce different levels of success. For this reason, science literacy analysis needs to examine student performance on each task, rather than relying solely on aggregate scores. Similar concerns have been raised in other content areas, where scientific literacy has been shown to be conceptualized differently across tasks embedded in junior secondary physics textbooks (Wan et al., 2024).

Task-level analysis is important because scientific literacy encompasses abilities that go beyond merely recalling facts and concepts. Students need to use scientific knowledge to explain phenomena, evaluate investigations, interpret evidence, and make decisions on issues related to science. The PISA framework positions these abilities as provisions for understanding science-related issues and making responsible decisions (OECD, 2016, 2025). The development of scientific literacy from the primary education level onward is therefore not only intended to support science learning, but also to equip students to understand and respond to everyday problems that involve science. This orientation aligns with the argument that embedding science assessment in real-world contexts is central to preparing students to use scientific knowledge in everyday situations (Fensham, 2009).

Scientific literacy, however, is a multidimensional construct whose performance is always manifested through particular tasks. In assessment, a competency such as explaining phenomena scientifically can be operationalized through tasks that ask students to identify explanations, make predictions, use representations, or apply scientific knowledge. Likewise, interpreting data can involve activities ranging from identifying simple patterns to evaluating conclusions based on a body of evidence (OECD, 2016). Consequently, the same competency label does not guarantee that every item imposes an identical demand. OECD (2016) explicitly distinguishes item difficulty, which is shown empirically by the proportion of participants who succeed on an item, from cognitive demand, the mental processes required to complete the task. This distinction is a reminder that a student's performance on a given item should not automatically be treated as a single representation of competency mastery. A comparable distinction underlies the task analysis framework that characterizes cognitive demand in terms of the integration of content and scientific practices required by a task (Tekkumru-Kisa et al., 2015).

Prior studies indicate that science literacy assessment has moved from measuring general achievement toward greater attention to instrument quality and item characteristics. Fives et al. (2014), for instance, developed a Scientific Literacy Assessment through a series of item analyses, think-aloud interviews, and expert judgment to ensure that the assessment tasks represented the intended conceptualization of scientific literacy. In the context of Indonesian primary schools, Muniroh et al. (2022) developed a science literacy assessment based on socio-scientific contexts and examined its validity, reliability, difficulty level, and item discrimination. More recent studies have also applied an Item Response Theory approach to examine the psychometric characteristics of science literacy instruments for primary school students (Nurhasanah et al., 2025). Meanwhile, the PISA assessment framework positions competency, knowledge type, and cognitive demand as dimensions that must be considered in item construction (OECD, 2016, 2025). This body of work reflects growing attention to the quality of science literacy measurement, yet most of this attention has been directed toward instrument development and validation or toward reporting achievement, while variation in performance across tasks—particularly tasks intended to represent the same competency—still leaves important room for analysis. In a related effort, items and scales for a science

inquiry test designed for primary education students have also been developed and validated (Pedaste et al., 2021). Item response theory has likewise been applied beyond science literacy, for example in the construction of a reading literacy item bank for fourth-grade students (Chen et al., 2023).

Attention to the task level is relevant because interpretation based solely on competency scores can produce an overly flattened picture. A given competency may show a certain average achievement, while the items that compose it produce very different levels of success. Such variation does not by itself prove that one task carries a higher cognitive demand than another, since empirical difficulty and cognitive demand are two different things (OECD, 2016). Nevertheless, this variation signals that the interpretation of science literacy performance needs to move beyond aggregate scores. Task-level analysis allows the identification of the specific parts of an assessment where student performance shifts, while also opening space to examine how scientific competencies are operationalized through different task characteristics. This perspective is pedagogically important because information that a student is "weak" or "strong" on a competency is not necessarily sufficient to indicate the type of learning experiences and science tasks that need to be strengthened. This is consistent with the argument that inquiry-based tasks should be evaluated against criteria that reflect authentic scientific epistemic practices rather than treated as generically interchangeable (Chinn & Malhotra, 2002).

Building on this issue, the present study positions task-level performance as an important unit of analysis for understanding the scientific literacy of primary school students. Unlike studies that are primarily oriented toward total scores, comparisons of competency achievement, or the testing of an instrument's psychometric characteristics, this study traces performance in a staged manner, from overall achievement, to achievement on the three competencies targeted by the instrument, to the pattern of success on each individual task. Particular focus is given to the variation among items within the same competency and to the conceptual demands apparent in each task. With this approach, the study aims to analyze variation in the science literacy performance of primary school students at the task level and to identify how this variation emerges both between and within science literacy competencies. This analysis is expected to provide a more differentiated interpretation of scientific literacy and to demonstrate the importance of considering not only the competency being measured, but also the specific task used to elicit that competency. Such an approach complements earlier Indonesian studies, including an analysis of prospective elementary teachers' initial science literacy and work on socio-scientific issue approaches to strengthen elementary students' scientific and cultural literacy (Widyasari & Haryanto, 2022). Related work has also examined socio-scientific issue approaches used to strengthen elementary students' scientific and cultural literacy (Wijaya et al., 2024), alongside the development of socio-scientific issue-based learning models aimed at strengthening elementary students' scientific literacy (Dessty et al., 2024).

Methodology

This study used a descriptive-exploratory design to analyze variation in students' scientific literacy performance at the competency and task levels. Initial data collection involved 402 students from three primary schools, comprising 108, 129, and 165 students, respectively. Data were obtained using a science literacy test consisting of 10 multiple-choice items with dichotomous scoring (1 = correct answer; 0 = incorrect answer). The instrument covered three science literacy competencies: explaining phenomena scientifically, represented by items 2, 3, 4, 7, and 9; interpreting data and evidence scientifically, represented by items 1, 5, and 8; and evaluating and designing scientific inquiry, represented by items 6 and 10. Prior to analysis, the data were checked to identify incomplete responses and values that did not conform to the 0–1 scoring scheme. The analysis used a *complete-case* approach, so that only responses that were complete and valid on all 10 items were included. Based on this process, 398 students were used as the analytic sample in this study. Item-level analytic approaches of

this kind echo the use of item response theory to examine students' self-regulated learning strategies (Uesaka et al., 2022).

Data analysis was carried out in stages at three levels. First, overall performance was analyzed based on the distribution of total scores, mean, median, and standard deviation. Second, performance at the competency level was calculated based on the proportion of correct responses across all items representing each competency. Proportions were used so that achievement on the three competencies could be compared descriptively even though the number of items composing each competency differed. Third, task-level analysis was conducted by calculating the proportion of correct responses on each item and comparing variation in performance between items, especially between items intended to represent the same competency. This quantitative analysis was complemented by a conceptual examination of each task based on the targeted competency, the form of the stimulus, the activity required of students, and the reasoning demands needed to determine the answer. The characteristics of each task were then juxtaposed with the empirical performance on that item to identify patterns of variation at the task level. The analysis was descriptive-exploratory in nature; therefore, differences in performance between items were not interpreted as evidence of a causal relationship between task characteristics and students' level of success. Although the present analysis relied on descriptive proportions rather than a formal psychometric model, tools such as the mirt package are available in the R environment for more advanced item response theory analyses of this kind (Chalmers, 2012).

Result and Discussion

Overall Scientific Literacy Performance

The analysis was conducted on 398 students who had complete and valid responses on all 10 items of the science literacy test. Students' total scores ranged from 1 to 10, with a mean of 6.89 (SD = 1.66) and a median of 7. The score distribution shows that most students were concentrated in the medium-to-high score range. A score of 7 was the most frequently obtained score ($n = 98$), followed by a score of 8 ($n = 88$) and a score of 6 ($n = 71$). A total of 55 students obtained a score of 9, and 11 students answered all items correctly. At the lower end of the distribution, 43 students obtained a score of 5 and 20 students obtained a score of 4, while the number of students with scores of 1–3 was relatively small.

Table 1. Distribution of Students' Scientific Literacy Scores

Total Score	n	Percentage (%)
1	1	0.3
2	6	1.5
3	5	1.3
4	20	5.0
5	43	10.8
6	71	17.8
7	98	24.6
8	88	22.1
9	55	13.8
10	11	2.8
Total	398	100.0

Although the total score distribution provides a general picture of students' science literacy performance, it does not yet indicate on which competencies and tasks students' successes or difficulties occurred. Therefore, the next analysis separates performance according to the three competencies represented in the instrument.

Performance Across Scientific Literacy Competencies

Performance at the competency level was calculated based on the proportion of correct responses across all items representing each competency. The results show that *interpreting data and evidence scientifically* had the highest proportion of correct responses, at 74.7%, followed by *evaluating and designing scientific enquiry* at 72.6%. Meanwhile, *explaining phenomena scientifically* showed a lower proportion of correct responses, at 64.0%.

Table 2. Students' Performance Across Scientific Literacy Competencies

Scientific literacy competency	Items	Number of items	Correct responses (%)
Explaining phenomena scientifically	2, 3, 4, 7, 9	5	64.0
Interpreting data and evidence scientifically	1, 5, 8	3	74.7
Evaluating and designing scientific enquiry	6, 10	2	72.6

The difference at the competency level spans a range of 10.7 percentage points. However, when the analysis was disaggregated to the level of individual tasks, a much larger range of performance was found. **Task-Level Variation in Scientific Literacy Performance**

Task-level analysis revealed substantial variation in the proportion of correct responses. Across the 10 items, the proportion of correct responses ranged from 32.4% to 89.7%, yielding a range of 57.3 percentage points.

Table 3. Item-Level Performance and Intended Scientific Literacy Competencies

Item	Intended competency	Correct responses (%)
1.	Interpreting data and evidence	88.7
2.	Explaining phenomena	71.4
3.	Explaining phenomena	73.6
4.	Explaining phenomena	53.0
5.	Interpreting data and evidence	69.6
6.	Evaluating/ designing enquiry	87.7
7.	Explaining phenomena	89.7
8.	Interpreting data and evidence	65.8
9.	Explaining phenomena	32.4
10.	Evaluating/ designing enquiry	57.5

Item 7 showed the highest proportion of correct responses (89.7%), followed by item 1 (88.7%) and I6 (87.7%). Conversely, item 9 had the lowest proportion of correct responses (32.4%), followed by item 4 (53.0%) and item 10 (57.5%). Thus, the difference in performance between tasks was greater than the difference apparent when the data were aggregated at the competency level.

Within-Competency Task Variation

Variation between tasks was also found when the analysis was restricted to items intended to represent the same competency. The largest variation was found within the *explaining phenomena scientifically* competency. The five items in this competency produced proportions of correct responses of 71.4% (I2), 73.6% (I3), 53.0% (I4), 89.7% (I7), and 32.4% (I9). The range of performance within this competency thus reached 57.3 percentage points.

In *interpreting data and evidence scientifically*, performance ranged from 65.8% on I8 to 88.7% on I1, a range of 22.9 percentage points. Meanwhile, the two tasks within *evaluating and designing scientific enquiry* also showed a fairly large difference: 87.7% of students answered I6 correctly, while 57.5% answered I10 correctly, a difference of 30.2 percentage points.

Table 4. Within-Competency Variation in Task Performance

Competency		Lowest-performing item	Highest-performing item	Min (%)	Max (%)	Range (percentage points)
Explaining scientifically	phenomena	I9	I7	32.4	89.7	57.3
Interpreting data and evidence scientifically		I8	I1	65.8	88.7	22.9
Evaluating and designing scientific enquiry		I10	I6	57.5	87.7	30.2

This finding shows that aggregation at the competency level reduces the variation apparent at the task level. This pattern is clearest in *explaining phenomena scientifically*: the competency average of 64.0% lies between two very different performances, namely 89.7% on I7 and 32.4% on I9.

Conceptual Characteristics of Tasks and Observed Performance

An examination of the task content shows that items placed within the same competency do not always require identical intellectual activity. Within explaining phenomena scientifically, for instance, I2 asked students to use the relationship between biotic and abiotic components to explain an ecosystem, whereas I3 asked students to predict the consequences of a change in one component of an ecological relationship. I7 asked students to determine the most likely cause of water turbidity based on the contextual information provided. In contrast, I9 asked students to compare several alternative actions and choose the most appropriate one to support the return of small living creatures to the school environment. For these four task forms, the proportions of correct responses were 71.4%, 73.6%, 89.7%, and 32.4%, respectively. I9 explicitly presented several alternative actions, ranging from adding potted plants to creating a small garden with open soil and a variety of flowering plants.

A similar pattern is seen in the two tasks intended to represent evaluating and designing scientific enquiry. I6 asked students to choose the appropriate investigative procedure to test the relationship between the number of plants and the surrounding temperature. This task was answered correctly by 87.7% of students. I10, answered correctly by 57.5% of students, asked students to reconsider a plan for building a water-catchment area after new information emerged that the chosen location was also used as a route to the canteen. Thus, I6 focused on selecting a design for collecting evidence, whereas I10 involved evaluating and adjusting a plan based on additional information and site-use constraints. Within the competency of interpreting data and evidence scientifically, I1 asked students to identify a consistent pattern from data on changes in river conditions, litter, and the presence of small animals, and produced a success rate of 88.7%. Meanwhile, I5 asked students to evaluate several conclusions based on data on road conditions and temperature, with a success rate of 69.6%. I8 asked students to integrate several pieces of contextual information about tiled surfaces, the number of trees, drainage conditions, and rainfall to determine the most appropriate explanation for the occurrence of waterlogging; this task was answered correctly by 65.8% of students. The content of these three tasks shows that items placed within the same competency operationalize that competency through different task forms.

Overall, the results show two different levels of variation. Aggregate performance across the three competencies falls within a relatively narrower range, namely 64.0%–74.7%, whereas performance across the 10 tasks falls within a range of 32.4%–89.7%. Variation between tasks also remains apparent when items are compared only with other items within the same competency. These results provide a basis for further examining the relationship between competency grouping, task characteristics, and the interpretation of students' science literacy performance. These patterns are consistent with prior evidence that contextual

framing can shape the difficulty of PISA science items (Le Hébel et al., 2017). Language demands embedded in test items have similarly been shown to influence international assessment outcomes (El Masri et al., 2016).

Discussion

The main finding of this study shows that the picture of science literacy performance changes as the analysis moves from the competency level to the task level. At the competency level, the proportion of correct responses fell within a relatively narrow range, namely 64.0%–74.7%. At the task level, however, this range widened substantially to 32.4%–89.7%. This pattern indicates that an aggregate score at the competency level can mask the variation in performance that emerges at the level of individual tasks. This finding is consistent with the conception of scientific literacy as a construct whose performance is related not only to competency, but also to the knowledge, context, and cognitive demand involved in a task (OECD, 2017). Accordingly, achievement on a given competency should not be interpreted as a uniform performance characteristic across all tasks placed within that competency.

The most striking variation was found within explaining phenomena scientifically. Although all five tasks were placed within the same competency, the proportion of correct responses ranged from 32.4% to 89.7%. This difference becomes important when task characteristics are considered. I7 asked students to select the cause-and-effect explanation most consistent with information about water turbidity, whereas I9 asked students to evaluate several alternative actions to support the return of small living creatures to the school environment. A similar pattern was found in evaluating and designing scientific enquiry: I6, which asked students to choose an investigation design, obtained 87.7% correct responses, whereas I10, which required students to reconsider a plan based on new information and site constraints, obtained 57.5%. This difference supports the importance of attending to how a competency is operationalized into a concrete task. The PISA framework itself notes that intended items can differ from the demands ultimately operationalized by the actual items (OECD, 2017).

Nevertheless, variation in the percentage of correct responses cannot be directly interpreted as a difference in cognitive demand. OECD (2016) distinguishes empirical item difficulty, reflected in the proportion of participants who succeed on an item, from cognitive demand, the type and depth of the mental processes required to complete a task. A task may be difficult because students have not yet mastered the required knowledge, while another task may involve complex reasoning processes but is nevertheless completed successfully by many students. The demand of an item can also be influenced by the complexity of the knowledge involved, students' familiarity with the content, the cognitive operations required, and the use of models or abstract ideas (OECD, 2016). For this reason, the performance differences between I7–I9 or I6–I10 in this study are more appropriately understood as evidence of task-level variation that warrants further examination, rather than as evidence that one task intrinsically carries a higher cognitive demand than another.

This finding also has implications for science literacy assessment practice. Assessment development should not merely ensure that a number of items have been labeled according to a particular competency, but should also examine how each item actually elicits that competency. Fives et al. (2014), for example, demonstrated the importance of combining analysis of student responses, think-aloud interviews, and expert judgment in developing science literacy assessments. Such an approach is important because the validity of score interpretation depends on the alignment between the conceptualization of scientific literacy and the tasks used to represent it. The findings of this study reinforce this need from the perspective of empirical performance: tasks placed within the same competency can produce very different patterns of success. Item-level analysis should therefore be an important part of interpreting assessment results, not merely a technical stage in instrument development. Test language has similarly been shown to influence PISA outcomes across countries, reinforcing the need to examine assessment items closely (Jerrim et al., 2022). Instructional approaches

such as inquiry-based teaching have also been linked to stronger PISA science performance across multiple countries (Oliver et al., 2021).

Pedagogically, these results suggest that information about the "weakest competency" may not be sufficient by itself to determine students' learning needs. Teachers can obtain more diagnostic information by attending to the types of tasks that students have and have not successfully completed. Students may be able to select an explanation supported by explicit information, yet struggle when required to evaluate several alternative actions, integrate multiple pieces of information, or adjust a decision to a particular condition. Science literacy instruction therefore needs to provide a variety of tasks that allow students to use scientific competencies across diverse contexts and demands. At the same time, because this study is descriptive-exploratory in nature and uses a limited number of items for each competency, the findings are not intended to establish a hierarchy of competency difficulty or a causal relationship between task characteristics and student performance. Its main contribution is to show that what appears, at the aggregate score, to be achievement on a competency can conceal important variation among the tasks that compose that score. Complementary instructional resources, such as picture storybooks designed to build elementary students' science literacy skills, offer one practical avenue for addressing this need (Rusydiana et al., 2023). Curriculum-level opportunities to embed socioscientific issues have likewise been identified as a means of strengthening literacy at the elementary level (Pessoa et al., 2025).

Conclusion

This study shows that the science literacy performance of the 398 students analyzed exhibited greater variation at the task level than at the competency level. The proportion of correct responses across the three competencies fell within a range of 64.0%–74.7%, whereas across the 10 individual tasks the range reached 32.4%–89.7%, including considerable variation among tasks intended to represent the same competency. This finding indicates that, within the sample and instrument used in this study, competency-level scores can mask the performance variation apparent at the task level. However, this difference does not by itself indicate that a task carries a higher or lower cognitive demand. Given that this study is descriptive-exploratory in nature, involves three schools, and uses a limited and unbalanced number of tasks for each competency, the findings are not intended to be broadly generalized. Within these limits, task-level analysis can complement the interpretation of aggregate scores and provide a more differentiated picture of students' science literacy performance. Broader curricular efforts, such as the integration of indigenous science knowledge to strengthen scientific literacy and environmental awareness, point to complementary directions for future elementary science education (Prasetyo et al., 2024). Innovations such as socio-scientific issue-based e-modules further illustrate the range of instructional strategies being explored to support students' scientific literacy development (Zarkasih et al., 2023).

References

- Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education*, 86(2), 175–218. <https://doi.org/10.1002/sce.10001>
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1–29. <https://doi.org/10.18637/jss.v048.i06>
- Chen, Q., Zheng, H., Fan, H., & Mo, L. (2023). Construction of a reading literacy test item bank for fourth graders based on item response theory. *Frontiers in Psychology*, 14, 1103853. <https://doi.org/10.3389/fpsyg.2023.1103853>
- Desstya, A., Sayekti, I. C., Rahmawati, F. P., Hidayati, Y. M., Kaltsum, H. U., Mahya, Y. M. Z., Kamila, S. J. P., Khotimah, N. A., Fidyastuti, N., & Diaz, A. M. (2024). Development of a natural and social sciences learning model based on socio-scientific issues to strengthen

- scientific literacy of Islamic elementary school students. *Jurnal Penelitian Pendidikan IPA*, 10(11), 8703–8711. <https://doi.org/10.29303/jppipa.v10i11.9094>
- El Masri, Y. H., Baird, J.-A., & Graesser, A. (2016). Language effects in international testing: The case of PISA 2006 science items. *Assessment in Education: Principles, Policy & Practice*, 23(4), 427–450. <https://doi.org/10.1080/0969594X.2016.1218323>
- Fensham, P. J. (2009). Real world contexts in PISA science: Implications for context-based science education. *Journal of Research in Science Teaching*, 46(8), 884–896. <https://doi.org/10.1002/tea.20334>
- Fives, H., Huebner, W., Birnbaum, A. S., & Nicolich, M. (2014). Developing a measure of scientific literacy for middle school students. *Science Education*, 98(4), 549–580. <https://doi.org/10.1002/sce.21115>
- Jerrim, J., López-Agudo, L. A., & Marcenaro-Gutiérrez, O. D. (2022). The impact of test language on PISA scores: New evidence from Wales. *British Educational Research Journal*, 48(3), 420–445. <https://doi.org/10.1002/berj.3774>
- Le Hébel, F., Montpied, P., & Tiberghien, A. (2017). Sources of difficulty in assessment: Example of PISA science items. *International Journal of Science Education*, 39(4), 468–487. <https://doi.org/10.1080/09500693.2017.1294784>
- Muniroh, N., Rusilowati, A., & Isaeni, W. (2022). Instrument development of science literacy assessment with socio-sciences contains in natural science learning for elementary school. *Journal of Research and Educational Research Evaluation*, 11(1). <https://doi.org/10.15294/jere.v11i1.55421>
- Nurhasanah, Astari, T., Cahyanita, E., & Nuriman. (2025). Psychometric properties of a local wisdom-based science literacy instrument for primary school students: An item response theory approach. *Jurnal Penelitian Pendidikan IPA*, 11(11). <https://doi.org/10.29303/jppipa.v11i11.12973>
- Organisation for Economic Co-operation and Development. (2016). *PISA 2015 assessment and analytical framework: Science, reading, mathematic and financial literacy*. OECD Publishing. <https://doi.org/10.1787/9789264255425-en>
- Organisation for Economic Co-operation and Development. (2017). *PISA 2015 assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving* (Rev. ed.). OECD Publishing. <https://doi.org/10.1787/9789264281820-en>
- Organisation for Economic Co-operation and Development. (2025). *PISA 2025 science framework*. OECD.
- Oliver, M., McConney, A., & Woods-McConney, A. (2021). The efficacy of inquiry-based instruction in science: A comparative analysis of six countries using PISA 2015. *Research in Science Education*, 51(Suppl. 2), 595–616. <https://doi.org/10.1007/s11165-019-09901-0>
- Pedaste, M., Baucal, A., & Reisenbuk, E. (2021). Towards a science inquiry test in primary education: Development of items and scales. *International Journal of STEM Education*, 8(1), Article 44. <https://doi.org/10.1186/s40594-021-00278-z>
- Pessoa, P., de Lima, J., Piacentini, V., Realdon, G., Jeffries, A., Ometto, L., Lopes, J. B., Zeidler, D. L., Fonseca, M. J., Sousa, B., Pinto, A., & Sá-Pinto, X. (2025). A framework to identify opportunities to address socioscientific issues in the elementary school curricula: A case study from England, Italy, and Portugal. *PLoS ONE*, 20(3), e0308901. <https://doi.org/10.1371/journal.pone.0308901>
- Prasetyo, Z. K., Syawaludin, A., & Dessty, A. (2024). Integration indigenous science in merdeka curriculum to strengthen scientific literacy and environmental care in elementary schools: Need analysis. *Edelweiss Applied Science and Technology*, 8(5), 1062–1070. <https://doi.org/10.55214/25768484.v8i5.1806>
- Rusydiana, U., Widodo, W., & Suprpto, N. (2023). The development of picture story book to improve the science literacy skills of grade 4 elementary school student. *Studies in Philosophy of Science and Education*, 4(1), 22–34. <https://doi.org/10.46627/sipose.v4i1.274>

- Tekkumru-Kisa, M., Stein, M. K., & Schunn, C. (2015). A framework for analyzing cognitive demand and content-practices integration: Task analysis guide in science. *Journal of Research in Science Teaching*, 52(5), 659–685. <https://doi.org/10.1002/tea.21208>
- Uesaka, Y., Suzuki, M., & Ichikawa, S. (2022). Analyzing students' learning strategies using item response theory: Toward assessment and instruction for self-regulated learning. *Frontiers in Education*, 7, 921844. <https://doi.org/10.3389/feduc.2022.921844>
- Wan, D., Lee, Y.-J., Chen, S., & Yu, Q. (2024). How scientific literacy is conceptualized in tasks from junior secondary physics textbooks. *International Journal of Science Education*. Advance online publication. <https://doi.org/10.1080/09500693.2024.2429045>
- Widyasari, A., & Haryanto, H. (2022). Analysis of students' initial scientific literacy of science in elementary school teacher education student. *Jurnal Inovasi Pendidikan IPA*, 8(1), 57–66. <https://doi.org/10.21831/jipi.v8i1.41667>
- Wijaya, I. K. W. B., Suryaningsih, N. M. A., Astiti, K. A., Poerwati, C. E., & Syahputra, A. (2024). Socio scientific issue (SSI) approach as a momentum to instill cultural literacy and scientific literacy through natural science learning. *Jurnal Pendidikan MIPA*, 25(2), 972–985. <https://doi.org/10.23960/jpmipa/v25i2.pp972-985>
- Zarkasih, Z., Diniya, D., Permana, N. D., & Kirana, I. (2023). Development of science e-modules based on socio-scientific issues integrated with the Quran verses to establish Pancasila student profiles. *Journal of Natural Science and Integration*, 6(1), 63–75. <https://doi.org/10.24014/jnsi.v6i1.21855>