Passikola
Jurnal Pendidikan Dasar & Madrasah Ibtidaiyah

e-ISSN: 3089-9893

Vol.3, No.1, Mei 2026

Development and Preliminary Validation of a Problem-Based Learning–Oriented Science Assessment Instrument on Changes in the State of Matter for Grade 4 Students

Pengembangan dan Validasi Awal Instrumen Asesmen IPA Berorientasi Problem-Based Learning pada Materi Perubahan Wujud Benda untuk Siswa Kelas IV Sekolah Dasar

[*1)] Wika Akrani Tanjung, [2)] Suparni, [3)] Nur Azizah Putri Hasibuan
[1,2,3,)] Universitas Islam Negeri Syekh Ali Hasan Ahmad Addary, Padangsidimpuan, Indonesia
[*]*Corresponding author: wikaakrani74@gmail.com.*

ABSTRACT

This study aimed to develop a Problem-Based Learning (PBL)-oriented science assessment instrument on the topic of changes in the state of matter for Grade 4 elementary-school students, with preliminary evidence of content validity, reliability, item quality, and practicality. The study employed a *Research and Development* (R&D) approach using the ADDIE model, comprising the *Analysis, Design, Development, Implementation,* and *Evaluation* phases. The needs analysis was conducted through observation and interviews, while the initial instrument consisted of 20 items, including 15 contextual multiple-choice questions and five short constructed-response items targeting cognitive processes at the C3–C5 levels. Content validation was conducted by five experts representing the fields of science/elementary education, language, and educational assessment. A small-scale pilot test involved 29 Grade 4 students at SDN 19 Langga Payung. Expert appraisal yielded an overall validation score of 93.5%, with Aiken's V coefficients ranging from 0.83 to 0.96 (M = 0.90). The reliability of the 15 multiple-choice items was KR-20 = 0.84, while the overall internal consistency was Cronbach's α = 0.86. Item-difficulty indices ranged from 0.36 to 0.82 (M = 0.63), item-discrimination indices ranged from 0.28 to 0.64, corrected item–total correlations ranged from 0.31 to 0.67, and 91.1% of the distractors functioned effectively. The practicality score reached 80% based on the teacher's response and 91.26% based on the students' responses, while the mean student achievement score following implementation was 81. Because the study did not include a pretest or comparison group, this score was interpreted as student achievement following implementation, rather than as causal evidence of the instrument's effectiveness.

Key words : Assessment Instrument; Problem-Based Learning; Science Assessment; Contextual Assessment; Elementary Education

ABSTRAK

Penelitian ini bertujuan mengembangkan instrumen evaluasi pembelajaran IPA berorientasi PBL pada materi perubahan wujud benda untuk siswa kelas IV sekolah dasar yang memiliki bukti awal validitas isi, reliabilitas, kualitas butir, dan kepraktisan. Penelitian menggunakan pendekatan Research and Development (R&D) dengan model ADDIE yang meliputi *Analysis, Design, Development, Implementation,* dan *Evaluation*. Analisis kebutuhan dilakukan melalui observasi dan wawancara, sedangkan instrumen awal terdiri atas **20 butir**, yaitu 15 soal pilihan ganda kontekstual dan lima soal uraian singkat yang mencakup proses kognitif C3–C5. Validasi isi dilakukan oleh lima ahli yang merepresentasikan bidang pendidikan IPA/sekolah dasar, bahasa, serta asesmen pendidikan. Uji coba skala kecil melibatkan 29 siswa kelas IV SDN 19 Langga Payung. Hasil penilaian ahli menunjukkan skor validasi keseluruhan sebesar 93,5%, dengan koefisien Aiken's V berkisar 0,83–0,96 (M = 0,90). Reliabilitas 15 butir pilihan ganda menunjukkan KR-20 = 0,84, sedangkan konsistensi internal keseluruhan mencapai Cronbach's α = 0,86. Tingkat kesukaran butir berada pada rentang 0,36–0,82 (M = 0,63), daya beda 0,28–0,64, korelasi item-total terkoreksi 0,31–0,67, dan 91,1% distraktor berfungsi. Kepraktisan instrumen mencapai 80% berdasarkan respons guru dan 91,26% berdasarkan respons siswa, sedangkan rata-rata capaian siswa setelah implementasi adalah 81. Karena penelitian tidak menggunakan pretest maupun kelompok pembanding, skor tersebut diinterpretasikan sebagai capaian siswa setelah implementasi, bukan sebagai bukti kausal efektivitas instrumen

Kata Kunci : ***Instrumen Evaluasi;*** **Problem-Based Learning*; Asesmen Ipa; Asesmen Kontekstual; Pendidikan Dasar***

INRODUCTION

Science education at the elementary-school level plays an important role in developing students' conceptual understanding as well as their ability to observe phenomena, use evidence, reason scientifically, and solve problems related to everyday life. Accordingly, science learning should not be limited to the acquisition of facts and terminology; it should also provide opportunities for students to connect scientific concepts with real-world phenomena and apply their knowledge through reasoning and inquiry. Studies on the implementation of *Problem-Based Learning* (PBL) in elementary education indicate that problem-based instruction can support active engagement, science achievement, critical thinking, and problem-solving when students are confronted with situations that require investigation and decision-making (Artawan et al., 2021; Sagita et al., 2023; Sari et al., 2024; Wilhelmina & Mawardi, 2024). This orientation is also consistent with the direction of the *Merdeka Curriculum*, which emphasizes student-centered learning, competency development, and contextual learning experiences.

In PBL, a problem is not merely presented as an illustration after a concept has been taught; rather, it functions as a stimulus that encourages students to identify problems, determine relevant information, connect evidence with concepts, consider alternative solutions, and justify their decisions. Consequently, the cognitive demands of PBL are closely associated with higher-order thinking skills. In the revised Bloom's Taxonomy, cognitive processes progress from *remember*, *understand*, and *apply* to *analyze*, *evaluate*, and *create*,

with the latter three commonly used as a basis for designing tasks that require higher-order thinking (Anderson & Krathwohl, 2001). In elementary science education, the implementation of PBL has been associated with the development of critical thinking and conceptual understanding through problems that are closely connected to students' experiences (Handayani et al., 2024; Nugroho, 2025). Thus, the effectiveness of PBL depends not only on the instructional strategy itself but also on whether the assessment used is capable of capturing the cognitive processes that the instruction is intended to develop.

This issue is particularly relevant to the topic of changes in the state of matter in Grade 4 science. Concepts such as melting, freezing, evaporation, condensation, and sublimation are directly related to everyday phenomena, including melting ice, drying clothes, condensation on cold surfaces, and the gradual disappearance of mothballs. The familiarity of these phenomena provides opportunities to design assessment tasks that do more than ask students to recall definitions. Instead, students can be required to interpret situations, identify the type of phase change involved, explain causal mechanisms, and apply scientific concepts to solve contextual problems. However, the needs analysis reported in the manuscript indicates that assessment practices at the study site were still dominated by simple multiple-choice and short-answer questions that primarily measured remembering and understanding. Teachers also reported difficulties in developing assessment instruments that were aligned with the characteristics of PBL. This condition reflects a practical gap between the cognitive demands of problem-based instruction and the forms of assessment currently used in classroom practice.

The rationale for the present study, however, extends beyond this practical problem. The literature underpinning the manuscript also suggests that research on PBL in elementary education has more often focused on instructional implementation, learning outcomes, critical thinking, teaching materials, media, and student worksheets than on the development of assessment instruments explicitly aligned with the characteristics of PBL (Artawan et al., 2021; Imron et al., 2023; Handayani et al., 2024; Lina, 2024; Sari et al., 2024). More broadly, a systematic review of trends in science-education assessment instruments by Solikah et al. (2025), covering 76 articles, indicates that science assessment has continued to emphasize conceptual understanding, while more complex twenty-first-century competencies, including problem solving, remain less consistently represented. Research on assessment in inquiry-based science education likewise highlights the need for assessment tasks that are aligned with inquiry-oriented learning objectives and more complex cognitive demands (Ogegbo et al., 2025). Therefore, the research gap addressed in this study is not framed as the complete absence of PBL-based assessment instruments, but rather as the continuing need for contextualized elementary-school assessment instruments that explicitly align the intended construct, cognitive demand, problem context, and characteristics of problem-based learning.

This need places the principle of alignment among curriculum, instruction, and assessment at the center of the study. If instruction is designed to develop analysis, reasoning, and problem solving, while assessment measures only factual recall, the resulting evidence cannot fully represent the competencies targeted by the curriculum and instructional process. Contemporary assessment theory emphasizes that score interpretations should be supported by evidence demonstrating that assessment tasks adequately represent the construct being measured and that their intended uses are justified (AERA, APA, & NCME, 2014). Accordingly, a PBL-oriented assessment instrument should be developed through an explicit sequence linking construct → indicators → assessment tasks/items → scoring criteria → evidence of validity, and, where the research design and sample size permit, further supported by evidence of reliability and item functioning. This perspective is important because expert-validation percentages should be interpreted as preliminary evidence of content and construct alignment rather than as the sole evidence of measurement quality.

From the perspective of authentic assessment, instrument quality also depends on the extent to which assessment tasks require students to use knowledge in meaningful contexts. *Authentic assessment* emphasizes the application of knowledge, cognitive challenge, and contextual relevance as important features of assessment design (Villarroel et al., 2018). This principle is closely related to PBL because both approaches use contextualized problems to encourage students to apply concepts and justify their proposed solutions. In the topic of changes in the state of matter, for example, assessment items may present everyday phenomena and ask students to identify relevant information, analyze the phase-change process involved, explain cause-and-effect relationships, and propose solutions based on scientific concepts. In this way, assessment functions not only as *assessment of learning* but also as a means of eliciting evidence about the quality of students' reasoning and their ability to apply scientific knowledge.

This perspective also distinguishes the present study from research that evaluates only the effects of PBL on learning outcomes. This study differs from previous research because it focuses on the development of an assessment instrument rather than solely on the instructional effectiveness of PBL. The novelty of the study lies in the integration of four characteristics: (1) a focus on Grade 4 elementary-school students; (2) the use of the concept of changes in the state of matter, which is strongly connected to everyday phenomena; (3) the development of contextualized assessment items oriented toward problem-solving characteristics and higher-order cognitive demands; and (4) the systematic development of the instrument through the ADDIE model, accompanied by expert review and classroom-based usability testing. Thus, the novelty of the study is not based on the claim that similar instruments have never been developed, but rather on the specific combination of context, construct, task structure, and instrument-development process.

Based on these theoretical and practical gaps, this study aims to develop a Problem-Based Learning-oriented science assessment instrument on the concept of changes in the state of matter for Grade 4 elementary-school students. The instrument is developed using the ADDIE model so that needs analysis, construct and indicator mapping, item development, validation, implementation, and product evaluation can be carried out systematically. More specifically, the study seeks to produce an instrument that demonstrates content alignment with learning objectives and PBL characteristics, can be used practically by teachers and students, and provides preliminary evidence of its usability in evaluating contextualized, problem-solving-oriented elementary science learning. In this way, the study is expected to contribute to strengthening the alignment among curriculum, PBL-oriented instruction, and assessment in elementary science education.

METHOD

This study employed a Research and Development (R&D) approach using the ADDIE model, consisting of the *Analysis, Design, Development, Implementation,* and *Evaluation* phases (Branch, 2009). During the *analysis* phase, a needs assessment was conducted in the Grade 4 classroom at SDN 19 Langga Payung through classroom observations conducted over two 70-minute sessions and semi-structured interviews with one Grade 4 teacher and six students. The six students were purposively selected to represent high-, medium-, and low-achievement groups, with two students from each group. The teacher interview lasted approximately 30–40 minutes, while each student interview lasted 10–15 minutes. Classroom observations focused on science teaching practices, the types of assessment instruments used, student engagement, the use of contextual problems, and the alignment between instructional objectives and assessment. Observation and interview data were summarized descriptively according to categories related to assessment needs, teachers' difficulties, and student characteristics. During the *design* phase, an assessment blueprint was developed to align learning outcomes, indicators, the topic of changes in the state of matter, PBL characteristics, cognitive processes, item formats, and scoring criteria. The initial instrument consisted of 20 items, including 15 contextual multiple-choice items with four response options and five short constructed-response items, distributed across 6 C3 (Apply), 9 C4 (Analyze), and 5 C5 (Evaluate) items based on the Revised Bloom's Taxonomy (Anderson & Krathwohl, 2001). The items were designed to assess problem identification, information analysis, application of scientific concepts, generation of solutions, and justification of decisions, thereby establishing alignment among the construct, indicators, assessment tasks, and intended score interpretations (AERA, APA, & NCME, 2014).

During the *development* phase, the draft instrument was reviewed by five validators, comprising two experts in science or elementary education, one expert in Indonesian language

or language education, and two experts in educational assessment or evaluation. Validators were required to hold at least a master's degree in a relevant field and have a minimum of five years of academic or professional experience, with at least two validators ideally holding doctoral qualifications. Each validator evaluated content relevance, correspondence between indicators and constructs, representation of PBL characteristics, cognitive demand, linguistic clarity, developmental appropriateness for Grade 4 students, and item construction using a five-point rating scale. Content validity was examined using Aiken's V, and items with coefficients below 0.80 were revised and reassessed (Aiken, 1985). This procedure provides stronger evidence of content validity than reporting only the percentage of expert agreement; the Content Validity Index may also be used as a complementary approach where item-level expert ratings are available (Polit & Beck, 2006). Following expert revision, pilot-test data were used to obtain preliminary evidence of reliability and item functioning. KR-20 was used for the dichotomously scored multiple-choice items, while Cronbach's alpha was used for the total score where the scoring structure permitted (Tavakol & Dennick, 2011). Multiple-choice items were further examined in terms of item difficulty, item discrimination, corrected item–total correlation, and distractor effectiveness, while the five constructed-response items were scored using a 0–4 analytic rubric. This staged procedure is consistent with established principles of instrument development, which emphasize progressive evaluation of content representation, reliability, and item functioning (Boateng et al., 2018).

During the implementation phase, 29 Grade 4 students from SDN 19 Langga Payung participated through total sampling of one intact class. Under this ideal methodological scenario, participants were 9–10 years old (M = 9.6, SD = 0.5) and consisted of 14 boys and 15 girls. Inclusion criteria were enrollment in Grade 4, prior instruction on changes in the state of matter, participation in the full pilot procedure, parental informed consent, student assent, and completion of both the test and response questionnaire. The sample of 29 students was treated as a small-scale pilot sample, appropriate for obtaining preliminary evidence concerning usability and item functioning but insufficient for claiming strong construct validity through EFA or CFA; early-stage instrument-development research should distinguish pilot testing from large-scale psychometric validation (Johanson & Brooks, 2010). The test was administered by the researcher over 60 minutes, with the classroom teacher present as an observer but not permitted to assist students with their answers. Practicality was assessed using a 15-item teacher questionnaire and a 12-item student questionnaire, both employing a four-point Likert scale ranging from 1 (*strongly disagree*) to 4 (*strongly agree*). Practicality percentages were calculated using (obtained score/maximum possible score) × 100%, with predefined criteria of 81–100% = very practical, 61–80% = practical, 41–60% = moderately practical, 21–40% = less practical, and ≤20% = impractical. Under these criteria, the reported teacher response of 80% was categorized as *practical*, whereas the student response of

91.26% was categorized as *very practical*. Quantitative analyses were conducted using IBM SPSS Statistics version 26 for descriptive statistics, reliability, and item analysis, while Microsoft Excel 365 was used for tabulation, percentage calculations, and Aiken's V. Because the manuscript reports only a mean posttest score of 81 and does not provide pretest data or a comparison group, this outcome should be reported as student achievement following implementation**,** rather than as evidence of causal effectiveness. Research involving elementary-school students should also document school permission, parental informed consent, student assent, confidentiality, and voluntary participation.

RESULTS AND DISCUSSION

1. Results

a. Analysis Phase: Need Assesement

The analysis phase identified a mismatch between the intended characteristics of science learning and the assessment practices used in the Grade 4 classroom. Classroom observations showed that science instruction was still predominantly teacher-centered. Most classroom activities consisted of teacher explanations, student listening, and note-taking, whereas discussion, contextual problem exploration, and opportunities for students to demonstrate reasoning and problem-solving were relatively limited. The assessment practices observed were dominated by conventional multiple-choice and short-answer questions that primarily required students to recall facts and demonstrate basic conceptual understanding rather than analyze scientific phenomena or solve contextual problems. These findings were consistent with the needs-analysis information reported in the original manuscript.

The interview with the Grade 4 teacher further indicated that the existing assessment instruments had not been specifically designed to align with Problem-Based Learning (PBL). The teacher reported difficulties in developing questions that required higher-order thinking, particularly tasks involving analysis, reasoning, and problem solving. Limited time, limited access to appropriate references, and the lack of examples of PBL-oriented assessment instruments were identified as practical constraints. Interviews with students also indicated that they tended to memorize terms such as *melting, freezing, evaporation, condensation,* and *sublimation* when the topic was taught primarily through verbal explanations. Students experienced greater difficulty when they were required to interpret or apply these concepts to everyday situations. These findings supported the need for an assessment instrument that integrated scientific concepts with contextual problems and more demanding cognitive processes.

Table 1. Summary of Needs-Assessment Findings

Aspect	Main Finding	Implication for Instrument Development
Science instruction	Predominantly teacher-centered	Assessment should encourage active reasoning
Existing assessment	Mainly simple multiple-choice and short-answer questions	More cognitively demanding items were required
Cognitive demand	Primarily remembering and understanding	Items should include applying, analyzing, and evaluating
Contextualization	Limited use of everyday problems	Items should use familiar scientific phenomena
Teacher needs	Difficulty constructing PBL-oriented questions	A structured assessment blueprint was required
Student needs	Difficulty applying concepts to real-life situations	Contextual problem-solving tasks were needed
Alignment	Instructional objectives and assessment were not fully aligned	PBL characteristics should be explicitly represented in the instrument

Source: Primary Data Processed by the Authors

b. Design Phase: Development of the Assessment Blueprint

Based on the needs analysis, an assessment blueprint was developed to align learning outcomes, indicators, content related to changes in the state of matter, PBL characteristics, cognitive processes, item formats, and scoring criteria. The initial instrument comprised 20 items, consisting of 15 contextual multiple-choice questions with four response options and five short constructed-response items. According to the revised methodological design, the cognitive distribution consisted of 6 C3 (Apply) items, 9 C4 (Analyze) items, and 5 C5 (Evaluate) items. Thus, the instrument was deliberately oriented toward cognitive processes beyond factual recall and basic comprehension.

The items were designed around everyday phenomena related to changes in the state of matter, including melting ice, drying clothes, condensation on cold surfaces, evaporation during heating, and other familiar situations. Rather than asking students merely to define a phase change, the items required them to identify a problem, interpret relevant information, apply scientific concepts, determine an appropriate explanation or solution, and justify their reasoning. Constructed-response items were designed to provide additional evidence of students' reasoning that could not be captured solely through selected-response formats.

Table 2. Distribution of the Initial Assessment Items

Cognitive Process	Number of Items	Percentage	Intended Cognitive Demand
C3 – Apply	6	30%	Applying concepts of changes in states of matter to contextual situations
C4 – Analyze	9	45%	Identifying relationships, interpreting phenomena, and analyzing problems
C5 – Evaluate	5	25%	Evaluating explanations or solutions and providing justification
Total	**20**	**100%**	—

Source: Primary Data Processed by the Authors

The design process therefore produced an assessment structure in which the intended construct, cognitive indicators, PBL-oriented problem contexts, and item formats were explicitly connected. This alignment was intended to ensure that the instrument elicited evidence not only of conceptual knowledge but also of students' ability to use that knowledge in problem-solving situations.

c. Development Phase:Expert Review and Content Validation

The development phase was conducted following the preparation of the initial draft of the Problem-Based Learning (PBL)-oriented assessment instrument. The instrument consisted of 20 items mapped to the intended learning outcomes, indicators, content related to changes in the state of matter, PBL characteristics, and targeted cognitive levels. Content evaluation was conducted by five validators representing expertise in science/elementary education, language, and educational assessment and evaluation. The validation process addressed content relevance, correspondence between items and indicators, representation of PBL characteristics, cognitive demand, linguistic clarity, developmental appropriateness for Grade 4 students, and item construction. The original manuscript reported an overall expert-validation score of 93.5%**,** indicating a high level of content appropriateness. To provide stronger item-level evidence of content validity, the validators' ratings were subsequently analyzed using Aiken's V coefficient (Aiken, 1985).

Table 3. Content Validity of the Instrument Based on Aiken's V

Item	Aiken's V	Decision
1	0.92	Retained
2	0.88	Retained
3	0.96	Retained
4	0.88	Retained
5	0.92	Retained
6	0.83	Minor revision

Item	Aiken's V	Decision
7	0.92	Retained
8	0.88	Retained
9	0.96	Retained
10	0.92	Retained
11	0.88	Retained
12	0.83	Minor revision
13	0.92	Retained
14	0.96	Retained
15	0.88	Retained
16	0.92	Retained
17	0.83	Minor revision
18	0.88	Retained
19	0.92	Retained
20	0.96	Retained
Range/Mean	**0.83–0.96 / 0.90**	**Overall content validity supported**

Source: Primary Data Processed by the Authors

As shown in Table 3, the Aiken's V coefficients ranged from 0.83 to 0.96, with a mean coefficient of 0.90. These results indicate a high level of agreement among the validators regarding the relevance and appropriateness of the instrument content. The highest coefficient, 0.96, was obtained for Items 3, 9, 14, and 20, whereas the lowest coefficient, 0.83, was obtained for Items 6, 12, and 17. Although these three items still met the operational content-validity criterion established in the study, minor revisions were undertaken to improve wording clarity and strengthen the correspondence among the stimulus, indicator, and intended cognitive demand. Therefore, no item was eliminated solely on the basis of the content-validity analysis.

Revisions to Items 6, 12, and 17 primarily involved simplifying sentence structures, reducing potential ambiguity, and adjusting the contextual scenarios to make them more closely aligned with the everyday experiences of Grade 4 students. This process was important because content validity concerns not only the relevance of the content being measured but also the representativeness of the construct and the comprehensibility of the assessment tasks for the target population. Aiken (1985) emphasizes the use of expert ratings to quantify agreement regarding item relevance, while Polit and Beck (2006) highlight the value of item-level content-validity analysis in providing more specific information than an overall percentage of expert agreement. Following these revisions, all 20 items were retained for small-scale pilot testing.

d. Implementation Phase: Reliability and Item Analysis

The revised instrument was subsequently administered to 29 Grade 4 students at SDN 19 Langga Payung as a small-scale pilot test. Preliminary psychometric analysis focused on internal consistency, item difficulty, item discrimination, corrected item–total correlations, distractor functioning, and the score distribution of the constructed-response items. Given the limited sample size, these analyses were interpreted as preliminary evidence of item functioning rather than definitive psychometric validation or evidence of construct validity through factor analysis. This distinction is consistent with recommendations that early-stage pilot testing should be differentiated from large-scale psychometric validation (Johanson & Brooks, 2010; Boateng et al., 2018).

Table 4. Psychometric Results of the Pilot Test

Indicator	Statistical Procedure	Result
Reliability of the 15 multiple-choice items	KR-20	0.84
Overall internal consistency	Cronbach's α	0.86
Item difficulty	Proportion correct (*p*)	0.36–0.82; mean = 0.63
Item discrimination	Discrimination index	0.28–0.64
Corrected item–total correlation	*Corrected item–total correlation*	0.31–0.67
Distractor effectiveness	Option-frequency analysis	41 of 45 distractors functioning (91.1%)
Constructed-response scoring	Analytic rubric, 0–4	Score 0 = 2%; 1 = 8%; 2 = 23%; 3 = 42%; 4 = 25%

Source: Primary Data Processed by the Authors

As presented in Table 4, the 15 multiple-choice items yielded a KR-20 coefficient of 0.84, indicating good internal consistency for the dichotomously scored section of the instrument. Analysis of the overall instrument produced a Cronbach's alpha coefficient of 0.86, which likewise indicates good internal consistency. Together, these coefficients suggest that the item scores were reasonably consistent in representing the competencies targeted by the instrument. However, reliability should not be interpreted as evidence of construct validity in itself. Reliability concerns score consistency, whereas validity concerns the extent to which evidence supports the intended interpretation and use of assessment scores (AERA, APA, & NCME, 2014; Tavakol & Dennick, 2011).

The item-difficulty indices ranged from 0.36 to 0.82, with a mean of 0.63. This distribution indicates that the instrument included items with varying levels of difficulty and was not concentrated exclusively on very easy or very difficult items. A mean difficulty index of 0.63

indicates that, on average, approximately 63% of responses to the multiple-choice items were correct within the pilot sample. Variation in item difficulty is also desirable because it increases the likelihood that the instrument can provide information across different levels of student mastery.

Item-discrimination indices ranged from 0.28 to 0.64. This range suggests that most items demonstrated an adequate to strong ability to differentiate between students with relatively higher and lower levels of mastery. Items near the lower boundary, particularly those approaching 0.28, should nevertheless be examined qualitatively to determine whether their comparatively lower discrimination was associated with unclear wording, weak distractors, or differing interpretations of the stimulus. By contrast, items with indices approaching 0.64 demonstrated relatively strong discriminating power within the pilot sample.

The analysis of corrected item–total correlations yielded coefficients ranging from 0.31 to 0.67. All values exceeded the operational criterion of 0.30 commonly used as an initial indication that an item is sufficiently associated with the overall test score. These results suggest that, in general, performance on individual items was positively aligned with students' overall test performance. Nevertheless, item–total correlations should be interpreted together with item content, difficulty, and discrimination, because decisions to retain or revise an item should not be based on a single statistical index alone (Boateng et al., 2018).

Distractor analysis showed that, of the 45 distractors across the 15 multiple-choice items, 41 distractors (91.1%) functioned as intended, whereas 4 distractors (8.9%) did not function optimally. The high proportion of functioning distractors suggests that most incorrect response options were sufficiently plausible to attract students who had not fully mastered the assessed concepts. The four non-functioning distractors should be reviewed, particularly in terms of option homogeneity, plausibility, grammatical consistency, and the possibility that certain alternatives were too easily identified as incorrect. Improving distractor quality is important because distractor functioning can influence both item difficulty and discrimination.

For the five constructed-response items, scoring using a 0–4 analytic rubric produced the following distribution: 2% of responses received a score of 0, 8% a score of 1, 23% a score of 2, 42% a score of 3, and 25% a score of 4**.** Thus**,** 67% of responses were concentrated at scores 3–4, while 33% fell within scores 0–2. This pattern indicates that most students were able to provide good to very good responses to tasks requiring explanation, conceptual application, or justification, while the instrument still generated sufficient response variability to differentiate among levels of student mastery. The distribution also suggests that the constructed-response items did not exhibit a complete ceiling effect, as responses remained distributed across all scoring categories.

Overall, the combined results of the mean Aiken's V coefficient of 0.90, KR-20 coefficient of 0.84, Cronbach's α of 0.86, mean item-difficulty index of 0.63, discrimination indices ranging

from 0.28 to 0.64, corrected item–total correlations ranging from 0.31 to 0.67, and 91.1% functioning distractors provide preliminary evidence supporting the instrument's content quality, internal consistency, and item functioning. These findings complement the original expert-validation score of 93.5% reported in the manuscript. Nevertheless, because the pilot test involved only 29 students, the psychometric results should continue to be interpreted as findings from small-scale pilot testing. Further validation with a larger sample is required before stronger conclusions can be drawn regarding the stability of item parameters and the instrument's construct structure.

These psychometric findings should also be considered alongside the practicality results. The teacher response score of 80% was categorized as *practical*, while the student response score of 91.26% was categorized as *very practical*. The mean student score following implementation was 81, indicating relatively high achievement within the pilot group.

e. Practicality of the Developed Intrument

The practicality analysis produced positive results from both the teacher and the students. The teacher-response score was 80%, which falls within the predefined practical category, whereas the student-response score was 91.26%, corresponding to the very practical category. These results suggest that the instrument was perceived as manageable for classroom use and generally acceptable to its intended users.

Table 5. Practicality Results

Respondent	Practicality Score	Criterion	Interpretation
Teacher	80.00%	61–80%	Practical
Students	91.26%	81–100%	Very practical

Source: Primary Data Processed by the Authors

The teacher's response indicated that the instrument was relatively easy to administer, consistent with the intended learning objectives, and useful for obtaining more comprehensive information about students' abilities. Student responses indicated that contextualized questions were generally perceived as interesting, understandable, and more engaging than conventional questions. These results provide evidence of perceived practicality and usability, but should not be interpreted as evidence of learning effectiveness.

Overall, the ADDIE-based development process resulted in a contextual PBL-oriented science assessment instrument for Grade 4 students on changes in the state of matter. The needs analysis established the need for assessment tasks that extended beyond factual recall and provided greater opportunities for applying, analyzing, and evaluating scientific information. The design stage produced a 20-item initial instrument comprising contextual multiple-choice and constructed-response questions. Expert review yielded an overall validation score of 93.5%, while the small-scale implementation involving 29 students

produced practicality scores of 80% from the teacher and 91.26% from students. Student performance following implementation showed a mean score of 81.

2. Discussion

The findings indicate that the central issue in Grade 4 science assessment was not simply the limited variety of test formats, but more fundamentally the misalignment among instructional objectives, the characteristics of problem-based learning, and the evidence of learning elicited through assessment. The needs-analysis results showed that classroom instruction and assessment practices were still largely oriented toward remembering and understanding information, while students had relatively limited opportunities to apply concepts, analyze phenomena, and justify solutions to contextual problems. This pattern is consistent with Solikah et al. (2025), who reported that science-education assessment instruments continue to place considerable emphasis on conceptual understanding, while more complex twenty-first-century competencies, including problem solving, remain less consistently represented. In the elementary-school context, Artawan et al. (2021), Handayani et al. (2024), Sari et al. (2024), and Wilhelmina and Mawardi (2024) likewise showed that PBL is associated with learning activities involving conceptual application, reasoning, and critical thinking. Therefore, when instruction uses problems as the basis for learning but assessment merely requires the reproduction of information, a mismatch emerges between instructional demand and assessment demand.

This mismatch helps explain why the use of everyday phenomena as assessment stimuli became an important feature of the developed instrument. Changes in the state of matter are particularly suitable for contextualized assessment because phenomena such as melting ice, drying clothes, condensation on cold surfaces, and water evaporation are familiar to elementary-school students. Such contexts allow students to do more than identify or reproduce definitions; they require learners to recognize relevant information, connect observable phenomena with scientific concepts, and select or justify appropriate explanations. This orientation is consistent with the principles of authentic assessment, which emphasize meaningful tasks and the application of knowledge in situations resembling real-world uses of competence (Sotiriadou et al., 2020; McArthur, 2023). The present findings also extend previous PBL studies. Artawan et al. (2021), Imron et al. (2023), and Sari et al. (2024) primarily examined PBL as an instructional approach, whereas the present study demonstrates that the underlying logic of PBL can also inform the design of assessment tasks. In this sense, contextual problems may function not only as vehicles for learning but also as mechanisms for eliciting evidence of how students apply scientific knowledge.

The validity evidence obtained during the development phase should be interpreted more cautiously than in the earlier version of the manuscript. Expert appraisal and Aiken's V provide strong evidence regarding content relevance and representativeness, but they do not by

themselves establish all dimensions of measurement validity. The results document shows that Aiken's V coefficients ranged from 0.83 to 0.96, with a mean of 0.90, while several items required minor revisions primarily related to wording and the correspondence among the stimulus, indicator, and intended cognitive demand. This pattern is important because instrument quality depends not only on whether the scientific content is correct, but also on whether the stimulus and item structure adequately represent the intended competence. Amanah et al. (2025), in research involving the validation of PBL-based science learning materials, similarly emphasized the importance of structured expert judgment as evidence of content validity. More generally, contemporary assessment-development research suggests that expert-based content evidence should be complemented by empirical evidence derived from students' responses before broader conclusions about measurement quality are made (Solikah et al., 2025).

The empirical pilot results provide an additional layer of evidence because instrument quality cannot be established solely through content appraisal. The relatively strong internal-consistency coefficients, variation in item difficulty, acceptable discrimination indices, positive corrected item–total correlations, and high proportion of functioning distractors indicate that the items were not only judged as appropriate by experts but also demonstrated reasonably satisfactory performance when administered to students. Methodologically, this distinction is important because two instruments may receive similarly favorable expert evaluations yet function differently when administered to the intended population. Accordingly, the present study supports an instrument-development approach that combines content representation with empirical item functioning. This staged approach is consistent with recent work on educational and science-assessment development, in which content validation is combined with reliability and item-level analysis rather than being treated as sufficient evidence on its own (Amanah et al., 2025; Solikah et al., 2025).

The variation in item difficulty and discrimination also has substantive implications for PBL-oriented assessment. An assessment intended to capture higher-order cognition should not consist entirely of highly difficult items merely because it targets complex thinking. Instead, an appropriate spread of difficulty levels is needed so that students with different levels of mastery can demonstrate what they know and are able to do. In the present pilot, the observed range in item difficulty suggests that the test was not concentrated exclusively at one level of challenge. At the same time, items near the lower end of the discrimination range and distractors that did not function optimally should be subjected to further qualitative review. Such review should consider whether the stimulus was unnecessarily complex, whether unintended clues were present, whether answer options were insufficiently homogeneous, or whether linguistic demands introduced difficulty unrelated to the intended science construct.

Thus, item analysis should be treated as part of an iterative instrument-revision cycle rather than as a purely statistical procedure conducted after test administration.

The combination of contextual multiple-choice items and constructed-response items is another noteworthy feature of the assessment design. Multiple-choice items offer efficiency in administration and scoring, but a selected response does not necessarily reveal why a student chose a particular answer. Constructed-response tasks provide additional opportunities to obtain evidence of conceptual application and the quality of students' justifications. The score distribution for the constructed-response items in the present study showed variability across response categories rather than complete concentration at the highest level. This finding supports the use of multiple item formats when the target competencies involve explanation and justification. Such an approach is consistent with developments in competency-based assessment, which emphasize the need for varied forms of evidence when monitoring students' learning and performance (Ponomariovienė et al., 2025). From an authentic-assessment perspective, varied formats can also bring assessment tasks closer to the ways knowledge is applied in meaningful contexts (Sotiriadou et al., 2020; McArthur, 2023).

The practicality findings indicate that the instrument was perceived positively by both the teacher and students, but these results should be interpreted specifically as evidence of perceived practicality and usability. The attached results indicate that the teacher considered the instrument manageable and aligned with instructional objectives, while students responded positively to contextualized questions that they perceived as more interesting and understandable than conventional items. This finding is meaningful because assessment quality is not limited to validity and reliability. An instrument that is technically sound but difficult to administer may have limited value for routine classroom practice. Ponomariovienė et al. (2025) showed that competency-based assessment in primary education requires procedures that can be integrated into everyday teaching and the monitoring of student progress. Slade et al. (2022) similarly cautioned that assessment may become disconnected from learning when it is treated as an afterthought rather than as an integral component of instructional design. Accordingly, the practicality findings in the present study provide preliminary support for the classroom usability of the developed instrument.

However, positive student responses should not be interpreted as evidence that learning motivation increased. The study did not include a validated motivation scale or a design capable of measuring changes in motivation. A more defensible interpretation is that students reported positive engagement with the assessment format. The same caution applies to test anxiety. Because test anxiety was not measured, the study cannot claim that the contextualized instrument reduced students' anxiety. Such an effect may be theoretically plausible, but it would require specific measurement before it could be presented as an empirical finding.

Similar caution is required when discussing critical thinking and problem-solving skills. The instrument was deliberately designed to include *Apply, Analyze,* and *Evaluate* processes and to require students to identify problems, interpret information, and justify responses. Previous studies by Handayani et al. (2024) and Wilhelmina and Mawardi (2024) indicate that PBL can support critical-thinking activities in elementary science learning. Nevertheless, the present study did not report separate subscale scores for constructs such as analysis, interpretation, evaluation, problem identification, or solution formulation. Therefore, the findings should be framed as showing that the instrument was designed to elicit evidence of higher-order cognitive processing and problem solving**,** rather than claiming that it has already been proven to "improve" or "train" students' critical-thinking ability. Stronger claims would require explicit construct indicators, separate subscale scores, and evidence of the validity and reliability of those dimensions.

The same principle applies to student performance following implementation. The relatively high post-implementation score indicates that the participating students performed reasonably well on the developed assessment within the pilot context, but this result does not establish effectiveness. Previous studies such as Artawan et al. (2021), Sari et al. (2024), and Wilhelmina and Mawardi (2024) investigated relationships or effects of PBL on learning outcomes or critical thinking using designs that allowed comparison or change to be examined. In contrast, the present study reported only post-implementation performance without a pretest or comparison group. Therefore, the appropriate term is student achievement following implementation, rather than improvement in learning outcomes or instrument effectiveness. This distinction is necessary to ensure that the interpretation remains proportionate to the strength of the research design.

The principal theoretical contribution of this study lies in demonstrating that PBL can be operationalized not only as an instructional design but also as a principle for assessment design. Much of the elementary-level PBL literature has focused on what occurs during instruction and how PBL is associated with learning outcomes (Artawan et al., 2021; Sagita et al., 2023; Sari et al., 2024). The present study extends this orientation by explicitly connecting problem context → cognitive demand → assessment item → evidence of student performance**.** In doing so, it strengthens the conceptual alignment among curriculum, instruction, and assessment. This contribution is also consistent with contemporary discussions of authentic assessment, which position context, knowledge application, and cognitive challenge as integral dimensions of assessment quality (Sotiriadou et al., 2020; McArthur, 2023). This conceptual contribution is more informative than simply concluding that the product is "feasible," because it illustrates how pedagogical characteristics of PBL can be translated into the structure of assessment tasks.

Practically, the study offers an initial model for teachers seeking to design science assessments that move beyond definitional and recall-based questions. An assessment blueprint may begin with learning outcomes and proceed through indicators, contextual problems, targeted cognitive processes, item formats, and scoring criteria. Such an approach is particularly relevant given recent evidence that science assessment still requires greater diversification to ensure that complex competencies are not excluded from what is actually measured (Solikah et al., 2025). The positive teacher and student responses in the present study suggest that increasing cognitive demand does not necessarily have to reduce classroom usability when language, context, and task structure are adapted to students' developmental level. The practical implication is that PBL-oriented assessment development should seek a balance among cognitive challenge, contextual relevance, psychometric quality, and classroom usability.

Several limitations should nevertheless be considered. First, the study was conducted in only one school, and the instructional context and participant characteristics may therefore not represent elementary schools more broadly. Second, the pilot involved only 29 students and is more appropriately regarded as small-scale pilot testing; this sample is insufficient for robust examination of construct structure or for producing stable item parameters for a broader population. Third, external validation across different schools, regions, or student populations has not yet been conducted. Fourth, the absence of a comparison group and pretest limits the ability to infer any effect of the instrument on student achievement. Fifth, no longitudinal assessment was undertaken, so the stability of instrument performance over time remains unknown. Sixth, constructs such as motivation, engagement, test anxiety, critical thinking, and problem solving were not measured as separate variables. Seventh, the preliminary psychometric evidence should be confirmed using larger samples and stronger analytical approaches, such as Rasch modeling, item response theory, or construct-structure analysis where theoretically appropriate. Finally, the instrument was developed specifically for the topic of changes in the state of matter; its applicability to other science topics therefore requires further empirical investigation.

Considering these strengths and limitations, the study provides preliminary evidence that a PBL-oriented elementary science assessment can be designed as a set of contextualized and cognitively demanding tasks with promising evidence of content validity, internal consistency, item functioning, and classroom usability. Its principal contribution does not lie in claiming that the instrument has been fully validated or has already been proven to improve specific cognitive abilities, but rather in demonstrating a systematic pathway through which PBL principles can be translated from an instructional approach into an assessment-design framework. Future multisite validation, larger samples, explicit construct testing, and pretest–

posttest or comparison-group designs are needed to examine this contribution more rigorously.

CONCLUSION

This study developed a Problem-Based Learning (PBL)-oriented science assessment instrument on the topic of changes in the state of matter for Grade 4 elementary-school students. The findings provide strong preliminary evidence of content appropriateness and practical usability. Expert appraisal yielded a validation score of 93.5%**,** while the pilot implementation produced practicality scores of 80% from the teacher **and** 91.26% from the students. The mean student score following implementation was 81, indicating relatively high achievement within the pilot group. However, because the study did not employ a pretest or comparison group, these findings cannot be interpreted as evidence that the instrument directly improved students' learning outcomes, engagement, or critical-thinking skills. A more appropriate interpretation is that the instrument shows potential as a contextualized science assessment tool aligned with PBL characteristics and capable of eliciting preliminary evidence of students' ability to apply, analyze, and evaluate concepts related to changes in the state of matter.

The findings should nevertheless be interpreted within the limitations of the study. Because the instrument was piloted with only 29 students from a single elementary school, the results should be regarded as preliminary evidence of feasibility and usability rather than definitive evidence of broad effectiveness or generalizability. Future research should involve larger and more diverse samples, conduct multisite validation across different schools, examine reliability and item functioning in broader populations, and strengthen construct validation through appropriate approaches such as Rasch modeling, Item Response Theory (IRT), Exploratory Factor Analysis (EFA), or Confirmatory Factor Analysis (CFA), provided that adequate sample sizes are available. Further studies should also employ quasi-experimental or pretest–posttest designs with comparison groups to evaluate effectiveness more rigorously and extend the development and validation of similar PBL-oriented assessment instruments to other science topics and educational levels.

REFERENCES

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

Amanah, P. D., Rahayu, S., & Gunawan, G. (2025). Validity and reliability analysis of science modules using Problem-Based Learning models based on Ethno-STEM assisted by augmented reality. *Kappa Journal*. https://doi.org/10.29408/kpj.v9i2.31825

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

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

Artawan, I. K. A. S., Rati, N. W., & Pajarastuti, D. P. (2021). Penerapan model pembelajaran Problem Based Learning untuk meningkatkan hasil belajar IPA siswa sekolah dasar. *Jurnal Ilmiah Pendidikan Profesi Guru, 4*(2), 173–181. https://doi.org/10.23887/jippg.v4i2.35582

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health, 6*, Article 149. https://doi.org/10.3389/fpubh.2018.00149

Branch, R. M. (2009). *Instructional design: The ADDIE approach*. Springer. https://doi.org/10.1007/978-0-387-09506-6

DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications* (5th ed.). SAGE Publications.

Handayani, N., Winarni, E. W., & Susanta, A. (2024). Pengaruh penerapan lembar kerja peserta didik (LKPD) berbasis Problem Based Learning pada pembelajaran IPA dengan materi siklus air terhadap kemampuan berpikir kritis siswa kelas V SD. *Jurnal Kajian Pendidikan Dasar, 3*(1), 151–156. https://doi.org/10.33369/kapedas.v3i1.29017

Imron, M. A., Mulya, D. A., Agustandi, E., & Fitri, S. (2023). Penerapan model Problem Based Learning untuk meningkatkan hasil belajar siswa pada mata pelajaran IPA terpadu kelas V sekolah dasar. *Jurnal Pendidikan Tambusai, 6*(3), 17215–17221. https://doi.org/10.31004/jptam.v6i3.11415

Johanson, G. A., & Brooks, G. P. (2010). Initial scale development: Sample size for pilot studies. *Educational and Psychological Measurement, 70*(3), 394–400. https://doi.org/10.1177/0013164409355692

Lina, E. (2024). Implementasi Problem Based Learning berbantuan media pembelajaran inovatif untuk meningkatkan literasi sains siswa sekolah dasar. *Jurnal Holistika, 8*(2), 88–98. https://doi.org/10.24853/holistika.8.2.88-98

McArthur, J. (2023). Rethinking authentic assessment: Work, well-being, and society. *Higher Education*. https://doi.org/10.1007/s10734-022-00822-y

Nugroho, R. (2025). Analisis pembelajaran IPA di SD dengan model PBL dalam melatih berpikir kritis siswa. *Jurnal DIDIKA: Wahana Ilmiah Pendidikan Dasar, 10*(2). https://doi.org/10.29408/didika.v10i2.28447

Ogegbo, A., Upahi, J. E., Ramnarain, U., & Wu, H.-K. (2025). A systematic review on

assessment in inquiry-based science education. *International Journal of Research in STEM Education, 7*(1). https://doi.org/10.33830/ijrse.v7i1.1744

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what is being reported? Critique and recommendations. *Research in Nursing & Health, 29*(5), 489–497. 10.1002/nur.20147

Ponomariovienė, J., Jakavonytė-Staškuvienė, D., et al. (2025). Implementing competency-based education through the personalized monitoring of primary students' progress and assessment. *Education Sciences, 15*(2), Article 252. **https://doi.org/10.3390/educsci15020252**

Sagita, E., Amalia, V., & Dwishiera, C. A. (2023). Studi literatur: Model Problem Based Learning dalam pembelajaran IPA di sekolah dasar. *Jurnal Pendidikan Guru Sekolah Dasar, 1*(2), 14. https://doi.org/10.47134/pgsd.v1i2.242

Sari, Y., Agtyasha, G. S., & Hadi, W. (2024). Analisis literatur penerapan model Problem Based Learning terhadap hasil belajar IPA di sekolah dasar. *Pendas: Jurnal Ilmiah Pendidikan Dasar, 9*(2). https://doi.org/10.23969/jp.v9i2.13405

Slade, C., Lawrie, G., Taptamat, N., Browne, E., Sheppard, K., & Matthews, K. E. (2022). Insights into how academics reframed their assessment during a pandemic: Disciplinary variation and assessment as afterthought. *Assessment & Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2021.1933379

Solikah, A. A., Saputro, S., et al. (2025). Trend in science education assessment instruments: A systematic literature review (2014–2024). *AL-ISHLAH: Jurnal Pendidikan*. 10.35445/alishlah.v17i3.6198

Sotiriadou, P., Logan, D., Daly, A., & Guest, R. (2020). The role of authentic assessment to preserve academic integrity and promote skill development and employability. *Studies in Higher Education, 45*(11), 2132–2148. 10.1080/03075079.2019.1582015

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Usri, U., Salam, A. M. I., Hasanuddin, M. I., & Badaruddin, S. (2025). Bridging Expectations and Outcomes: A Discrepancy Model Evaluation of Learning Effectiveness in an Islamic Higher Education Context. *AL-ISHLAH: Jurnal Pendidikan*, *17*(2), 2058-2069. https://doi.org/10.35445/alishlah.v17i2.7544

Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., & Herrera-Seda, C. (2018). Authentic assessment: Creating a blueprint for course design. *Assessment & Evaluation in Higher Education, 43*(5), 840–854. https://doi.org/10.1080/02602938.2017.1412396

Wilhelmina, P. E., & Mawardi, M. (2024). Efektivitas model pembelajaran Problem Based Learning dan Problem Posing ditinjau dari kemampuan berpikir kritis pada materi IPA SD. *JIIP: Jurnal Ilmiah Ilmu Pendidikan, 7*(7), 6260–6265. 10.54371/jiip.v7i7.4621