



A DIGITAL SYSTEM-BASED ASSESSMENT FOR COLLABORATIVE PROBLEM-SOLVING SKILLS AMONG PRE-SERVICE SCIENCE TEACHERS: PRELIMINARY EVIDENCE OF QUALITY AND IMPLEMENTATION

W. B. Sabtiawan^{*1}, A. N. M. Fauziah¹, Wasis¹, T. Rijanto¹,
K. Wong-akharawin², A. F. Hendratmoko¹

¹Universitas Negeri Surabaya, Surabaya, Indonesia

²Yala Rajabhat University, Thailand

DOI: 10.15294/jpii.v15i3.54201

Accepted: June 12th, 2026. Approved: September 29th, 2026. Published: September 29th, 2026

ABSTRACT

Collaborative problem-solving (CPS) is central to science teacher preparation, yet existing assessments often emphasize individual outcomes or final products and provide limited evidence of how CPS develops across iterative stages of group problem-solving. This study developed and preliminarily evaluated a digital assessment system for science teachers' CPS across four stages: identifying and representing problems, planning and developing solutions, testing and refining solutions, and communicating evidence-based results. A multiphase, quantitative-dominant mixed-methods design involved three expert validators and 118 prospective science teachers from four intact classes. Quantitative evidence included content relevance (Aiken's V; expanded Gregory), internal structure (CFA), expert-judgment consistency (ICC), and multiclass score patterns; expert comments, revision records, and user feedback provided qualitative evidence integrated through a joint display. Content relevance was mixed, with CFA showing three dimensions loading above 0.30 and communicating results loading at 0.286. Expert-judgment consistency was inadequate, and student-score reliability was not established. Class scores differed significantly; however, intact-class observations preclude causal inference. Thus, the system provides preliminary stage-specific CPS score patterns rather than dependable individual-level profiles. Its contribution is integrating stage-specific assessment across an authentic design–test–redesign–retest cycle using analytic-rubric scoring with measurement and implementation evidence. Further refinement should prioritize communication descriptors, validator calibration, and student-score reliability.

© 2026 Science Education Study Program FMIPA UNNES Semarang

Keywords: collaborative problem-solving; digital assessment; quality science education

INTRODUCTION

Science education must develop critical, creative, communicative, and collaborative capabilities for solving complex problems (Fauziah et al., 2024; Hendratmoko et al., 2024; Nurita et al., 2025; Widodo et al., 2026). This priority aligns with Sustainable Development Goal 4, particularly its emphasis on developing relevant competencies and preparing qualified teachers who

can support effective learning. In science-teacher education, however, advancing these goals requires not only opportunities for collaboration but also assessment systems that generate evidence of how prospective teachers perform across different components of collaborative problem-solving (CPS). Such evidence can help teacher educators identify strengths and difficulties in collaborative performance, provide more targeted instructional feedback, and make better-informed decisions about designing and evaluating learning activities. In this way, improved CPS measurement

^{*}Correspondence Address

E-mail: wahyusabtiawan@unesa.ac.id

can contribute to quality education indirectly by strengthening the preparation and assessment of prospective science teachers rather than by serving as an outcome of SDG 4 in itself. PISA 2015 underscores the need for such preparation: only 7.9% of students across OECD systems reached the highest CPS proficiency level, while 28.1% performed below the baseline (OECD, 2017). CPS requires learners to understand problems, exchange information, construct shared representations, design and evaluate solutions, and coordinate roles (Sun et al., 2020; Tang et al., 2020; Thornhill-Miller et al., 2023). These processes are especially relevant to science learning, where investigation, argumentation, evidence evaluation, and collective decision-making often require learners to coordinate cognitive and social activity (Chen et al., 2020; Rosen et al., 2020; Negrón et al., 2026).

CPS is particularly important for prospective science teachers, who must master science while facilitating interactive, collaborative, and problem-based learning (Magaji, 2021; Anchunda & Kaewurai, 2025; Hariyono et al., 2026). They must manage discussion, promote equitable participation, guide problem identification, and direct inquiry toward scientifically defensible solutions (García-Carmona, 2020; Gillies, 2023; Siswanto et al., 2025). Science-education research has developed resources for integrating the 4Cs, collaborative problem-solving models, and measurement models for prospective teachers (Haryani et al., 2021; Sajidan et al., 2022; Hariyono et al., 2026), but assessment evidence about how prospective teachers collaborate while solving scientific problems remains limited.

Conceptually, CPS combines cognitive processes (understanding problems, planning, testing, and evaluating) with social processes such as shared understanding, role coordination, evidence exchange, joint monitoring, feedback, and strategy adaptation. The OECD (2017) framework similarly emphasizes perspectives, roles, shared representations, plans, monitoring, evaluation, and team adaptation. CPS is therefore a multidimensional performance construct that a conventional cognitive test or a single group-product score cannot adequately represent.

Preliminary observations and the assessment-document review conducted before development identified recurring omissions in observable criteria for role coordination, evidence exchange, joint monitoring, strategy adaptation, individual contribution, and collective communication. Existing practices documented concept mastery or final products more consistently than

the collaborative process. This practical gap supports earlier reports that lecturers rarely assess idea exchange, negotiation, coordination, and group-strategy evaluation systematically (Alfin et al., 2019; Meijer et al., 2020).

Existing digital CPS assessments differ substantially in their target populations, assessment contexts, evidence sources, scoring approaches, and reported measurement evidence. PISA 2015 assessed 15-year-old students through human-to-agent interactions in which students selected predefined communications and actions while working with computer-simulated partners; the resulting responses were scored at the individual level within a standardized psychometric framework (OECD, 2017; Rosen et al., 2020). The ATC21S assessment, in contrast, employed human-to-human dyadic tasks and captured time-stamped actions and communication through log-stream data, from which behavioral indicators of cognitive and social CPS skills were coded and modeled psychometrically (Care et al., 2016; Scoular & Care, 2019). The ETS Collaborative Science Assessment Prototype (ECSAP) moved closer to a disciplinary context by embedding CPS in science-oriented simulation tasks and capturing both collaboration processes and outcomes, including text-based interaction; however, its primary purpose was to explore standardized CPS assessment rather than assessment embedded in authentic science-teacher preparation (von Davier et al., 2017; Hao et al., 2019). More broadly, computer-based CPS assessments vary considerably in task design, participant age, interaction mode, evidence captured, and scoring procedures, and reliability and validity evidence remains incomplete for many available instruments (Chai et al., 2024).

These developments provide an important foundation for digital CPS measurement, but they remain insufficiently connected to authentic science-teacher education. Science-education studies have strengthened collaboration and problem solving through instructional interventions (García-Carmona, 2020; Haryani et al., 2021; Sajidan et al., 2022; Siswanto et al., 2025; Hariyono et al., 2026), but generally emphasize learning processes and outcomes rather than a dedicated system for assessing CPS performance across successive stages of authentic project work. Assessment and digital-learning studies, meanwhile, have emphasized indicators, analytic rubrics, validity evidence, data management, performance profiles, and feedback (Adom et al., 2020; Kaspi & Venkatraman, 2023; Wilson, 2023; Lin et al., 2024; Mutmainna et al., 2024; Hendratmoko et

al., 2026). Digital systems can further improve accessibility, traceability, data storage, and feedback processes (Maier & Klotz, 2022; Nugraha, 2022; Alenezi et al., 2023; Tomar et al., 2024), but digitalization alone does not ensure meaningful assessment; score interpretation still depends on the alignment of tasks, performance descriptors, scoring procedures, and appropriate validity evidence (Almanasreh et al., 2022; Brod et al., 2023; Carrillo-Avalos et al., 2025). The remaining gap is therefore not the absence of digital CPS assessment itself, but the limited integration of stage-specific CPS assessment with authentic project performance in prospective science-teacher education. The present study addresses this gap by developing a digital system that uses analytic-rubric evidence to assess cognitive and collaborative performance across an iterative identifying–planning–testing–communicating cycle while examining both preliminary measurement evidence and classroom implementation. Its contribution is therefore primarily contextual and integrative rather than evidence of psychometric superiority, particularly because student-score reliability has yet to be established.

Addressing this gap requires more than transferring a conventional rubric to a digital platform. The assessment task must elicit the intended cognitive and social processes, the rubric must differentiate observable performance levels, and the system must preserve traceable evidence for scoring and feedback. Accordingly, technological functionality and measurement quality are evaluated as interconnected features rather than as separate development outcomes.

This study therefore developed and preliminarily evaluated a digital CPS assessment system for prospective science teachers. The system integrates four connected CPS dimensions: (1) identifying and representing the problem, (2) planning and developing a solution, (3) testing and refining the solution, and (4) communicating evidence-based results, with observable collaborative performance, digital analytic-rubric scoring, and complementary measurement and implementation evidence. Rather than positioning this integration as psychometric superiority over existing CPS assessments, the study examines its feasibility and preliminary measurement quality within authentic science-teacher education.

Accordingly, the study addressed the following research questions: (RQ1) How was the digital CPS assessment system, including its task, analytic rubric, and digital workflow, developed and refined through expert review? (RQ2) To what extent did the assessment components

demonstrate content relevance based on Aiken's V, the expanded Gregory method, and expert qualitative judgments? (RQ3) What preliminary evidence supported the internal structure of the four-dimensional CPS model and the consistency of expert judgments, as examined through confirmatory factor analysis (CFA) and intraclass correlation coefficients (ICC)? (RQ4) What CPS score patterns and implementation characteristics emerged across the four intact classes, based on descriptive statistics, between-class comparisons, user responses, and qualitative implementation evidence? Together, these questions examine the system's development, preliminary measurement evidence, and classroom applicability while maintaining appropriate caution in score interpretation because student-score reliability has not yet been established.

METHODS

This study employed a multiphase, quantitative-dominant mixed-methods design as its primary methodological framework. Within this framework, the development of the digital CPS assessment system followed a Type 1 design and development research orientation, which focuses on designing, developing, and evaluating a specific educational product (Richey & Klein, 2007). The study comprised three connected phases. Phase 1 focused on developing and refining the authentic CPS task, analytic rubric, and digital workflow through expert review, combining quantitative content-relevance indices with qualitative expert comments and revision records. Phase 2 examined preliminary measurement evidence, including the internal structure of the four CPS dimensions and the consistency of expert judgments. Phase 3 examined multiclass implementation through CPS score profiles, between-class comparisons, user responses, and qualitative implementation evidence.

The quantitative strand was dominant because the primary evaluative claims relied on numerical evidence on content relevance, internal structure, expert-judgment consistency, CPS score patterns, and user responses. The qualitative strand was necessary because these quantitative indices alone could not explain why particular assessment components required revision, how experts interpreted weaknesses in the task and rubric, or how the system functioned during classroom implementation. Integration therefore occurred at two points. During development, expert comments and revision records were linked directly to quantitative content-relevance evidence

to guide revisions to the assessment components. At the interpretation stage, the study formally integrated quantitative and qualitative findings through a joint display to identify convergence, complementarity, and discrepancies across measurement and implementation evidence (Fetters et al., 2013; Creswell & Plato Clark, 2018). This design allowed the study to evaluate not only whether the digital CPS assessment system produced favorable quantitative indicators, but also how to interpret those indicators in relation to design revisions and classroom applicability.

The generic product-and-tool development process was modified into four sequential and iterative stages: (1) preliminary analysis and construct specification, (2) assessment blueprint and digital-prototype development, (3) expert validation and iterative refinement, and (4) empirical validation and multiclass implementation. This modification connected each design decision to evidence, analysis, and revision. Figure 1 summarizes the framework.

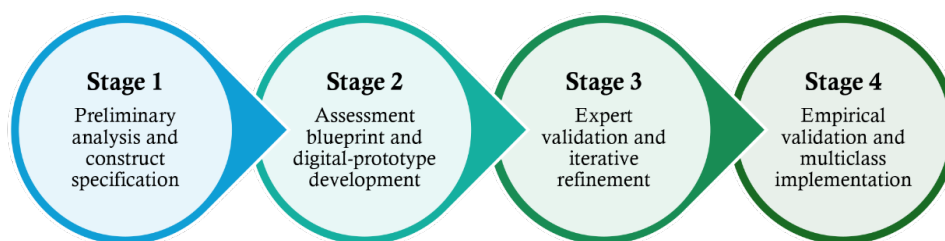


Figure 1. Modified Development Stages

The study was conducted in the Science Education Study Program of a state university in Indonesia during the second semester of the 2025/2026 academic year. Participants included three expert validators and 118 first-year prospective science teachers enrolled in the General Physics course across four intact classes: Class A ($n = 24$), Class B ($n = 32$), Class C ($n = 31$), and Class D ($n = 31$). The expert validators were purposively selected based on their expertise in science education, educational assessment and evaluation, learning technology, and/or collaborative problem-solving. Student participants were included if they were actively enrolled in the General Physics course, participated in all required CPS assessment activities, and had complete assessment data; students who did not complete all stages of the study or had incomplete assessment scores were excluded. The sample size was also evaluated for adequacy for the planned confirmatory factor analysis. With four observed indicators specified as loading on a single latent factor, the model estimated eight free parameters (four standardized loadings and four indicator residual variances, with the factor variance fixed for identification), yielding a participant-to-parameter ratio of approximately 14.75:1 for $N = 118$. This ratio exceeds the commonly recommended minimum 10:1 for simple, well-defined CFA models and approaches the general $N = 100$ –150 guideline for models with few indicators (Kline, 2023). Because the model nonetheless has few indicators and correspondingly low degrees of freedom, N

$= 118$ was judged minimally adequate for a preliminary one-factor CFA rather than sufficient for more complex or multigroup structural analyses. We did not assign the four classes to experimental or control conditions; they served as naturally occurring implementation contexts. Accordingly, we conducted between-class analyses to examine variation in individual CPS score patterns across these contexts and interpreted them descriptively rather than as evidence of treatment effects or causal class differences.

The research product combined contextualized science-project tasks, a four-level analytic rubric, and a digital system designed to elicit, record, and score CPS performance. The four related dimensions were (I1) collaboratively identifying and representing the problem, (I2) jointly planning and developing a solution, (I3) collaboratively testing, evaluating, and refining the solution, and (I4) communicating evidence-based results. The performance descriptors embedded shared understanding, participation, role coordination, evidence exchange, joint monitoring, and collective decision-making.

Students worked in groups of four to five across eight sessions to engage in the CPS process. Shared group evidence included problem representations, design plans, prototypes, testing records, and final group products. These artifacts documented the collaborative problem-solving process but were not directly assigned as individual CPS scores. Following the group-based CPS activities, each student completed an individual

CPS test designed to assess the four CPS dimensions. The individual CPS scores used in the between-class and dimension-level analyses were derived from these individual test responses rather than the shared group products. The research team scored the responses using the developed analytic rubric, while the digital system facilitated task delivery, response submission, rubric-based scoring, score storage, CPS-profile visualization, and feedback documentation. Although the analytic scores were generated from individual CPS test responses, students had previously participated in collaborative group activities and were also nested within four intact classes. Accordingly, we treated the scores as participant-level outcomes, while considering potential dependence associated with shared group and class contexts when interpreting between-class comparisons. Because the analyses did not explicitly model this multilevel structure, we treated the one-way ANOVA as preliminary and acknowledged possible within-group and within-class dependence as a methodological limitation.

We assessed user perceptions of the digital CPS assessment system using a researcher-developed questionnaire with 20 items across four domains: ease of use, assessment effectiveness, support for collaboration and problem-solving, and motivation and satisfaction. Items were rated on a four-point Likert-type scale ranging from 1 to 4. The questionnaire was developed from relevant literature, the digital assessment system's characteristics and intended functions, and expert judgment to align items with the user experience aspects being evaluated. Content validity was examined through expert judgment and yielded an Aiken's V coefficient of 0.954, indicating strong content relevance; internal-consistency reliability was examined using Cronbach's alpha and yielded $\alpha = 0.956$, indicating excellent internal consistency. The questionnaire was administered online to all 118 participants after they completed the digital CPS assessment activities. Scores for each domain were converted to percentages by dividing the obtained score by the maximum possible score and multiplying by 100; the overall response percentage was calculated from the total obtained score across all 20 items relative to the maximum possible total score.

Stage 1: Preliminary Analysis and Construct Specification

The preliminary study provided an empirical basis for the research problem and the intended interpretation of CPS scores. Observations and assessment-document reviews examined

whether existing practices captured shared problem representation, participation, role coordination, evidence exchange, joint monitoring, strategy adaptation, individual contribution, and collective communication.

Each observed or documented feature was coded as present, partly present, or absent. Frequencies and percentages summarized the representation of CPS processes, while directed content analysis organized field notes and document evidence according to the predefined indicators. The findings were converted into a construct map specifying intended score use, four CPS dimensions, observable evidence, task requirements, rubric criteria, and required digital functions.

Stage 2: Assessment Blueprint and Digital-Prototype Development

The construct map guided an assessment blueprint linking each CPS dimension to contextualized science problems, expected evidence, four-level performance descriptors, and digital-system functions. Tasks required students to understand and represent a problem, formulate and develop a shared solution, test and refine the solution using evidence, and communicate the results. We checked internal alignment across the construct, instructions, expected evidence, rubric, and system workflow. A technical walkthrough examined navigation, response recording, scoring, data storage, and feedback before expert validation.

Stage 3: Expert Validation and Iterative Refinement

The three validators independently evaluated the four principal CPS components involving identifying and representing the problem, planning and developing a solution, testing and refining the solution, and communicating evidence-based results using a four-point relevance scale. The validation also covered the associated assessment indicators, tasks, rubric descriptors, language, and digital-system components. In total, the validators rated six task-level subcomponents distributed across the four CPS dimensions involving two subcomponents for identifying and representing the problem (individual and group problem identification), one subcomponent for planning and developing a solution, two subcomponents for testing and refining the solution (testing and evaluation; redesign and retesting), and one subcomponent for communicating evidence-based results, in addition to the associated rubric descriptors, task language, and digital-system components described above. We examined

quantitative content-relevance evidence using Aiken's *V* and the expanded Gregory method, while written comments addressed construct coverage, clarity, task authenticity, differentiation among performance levels, scoring feasibility, and system usability. Given the small number of validators, Aiken's *V* was interpreted cautiously using the exact right-tail probability rationale proposed by Aiken (1985), rather than relying solely on a conventional descriptive cutoff. With three validators and four response categories, $V = 1.00$ represents unanimous maximum relevance and meets the conventional exact $p < .05$ criterion, whereas lower coefficients were treated as descriptive evidence requiring corroboration from qualitative expert judgments. We integrated quantitative indices and qualitative comments in a revision matrix, so decisions to retain, revise, or remove components were based on converging evidence rather than a single content-validity coefficient.

Content relevance was estimated using Aiken's *V*: $(V = \sum s) / (n(c-1))$ where $s = r - l_0$, r is the validator's rating, l_0 is the lowest rating category, c is the number of rating categories, and n is the number of validators (Aiken, 1985). Values closer to 1 indicate stronger content relevance. We revised components with lower coefficients or substantive expert criticism rather than judging them solely through a fixed pass-fail threshold.

The expanded Gregory coefficient complemented Aiken's *V* to describe unanimous strong relevance among three validators. We classified ratings of 1–2 as weak and 3–4 as strong, producing eight possible rating patterns. We calculated the coefficient as $VG = H / (A + B + C + D + E + F + G + H)$, where H represents components rated strongly relevant by all three validators (Retnawati, 2016; Sincovich et al., 2019).

We examined consistency among expert judgments descriptively alongside the content-relevance indices and qualitative comments. Although an earlier statistical output included intraclass correlation coefficient (ICC) estimates, the original analysis specification did not retain sufficient information on the ICC effects model, agreement definition, measurement unit, confidence interval, or significance test. Because these specifications are necessary for reproducible interpretation of ICC estimates (Koo & Li, 2016), we did not use the incomplete ICC output as formal reliability evidence in the present study. We therefore interpreted expert-judgment evidence primarily through Aiken's *V*, the expanded Gregory method, individual rating patterns, and qualitative comments. We treated student CPS scores

as a separate measurement issue and did not infer reliability from expert-validation data.

Stage 4: Empirical Validation and Multiclass Implementation

After expert revision, we implemented the system with 118 prospective science teachers in four intact classes. Students completed the CPS tasks through the digital system, and the validated rubric scored their performance. The empirical analysis examined the internal structure of the four CPS dimensions and described score profiles and implementation consistency across authentic classroom contexts.

Construct validity is necessary to ensure that measurement results accurately reflect the abilities being measured. In instrument development, theory-based validation and content validity are insufficient to ensure a good fit between the construct and the empirical data. Therefore, CFA confirms the instrument's latent structure and examines the relationships between indicators and underlying factors (MacCallum & Austin, 2000; Brown & Moore, 2012). CFA also allows statistical evaluation of model fit using various goodness-of-fit indices to assess how well the hypothesized measurement model fits the data (Brown, 2015; Collier, 2026).

CFA analysis was conducted using JASP version 0.12. Model fit was evaluated using several goodness-of-fit indices: Chi-square/df (χ^2/df), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), Goodness of Fit Index (GFI), and Standardized Root Mean Square Residual (SRMR). A model is considered fit if it meets the recommended criteria: $\chi^2/df < 3$ (ideally < 2) and CFI, TLI, and GFI > 0.90 , with RMSEA and SRMR < 0.08 (Hidayat et al., 2019; Collier, 2026).

In addition to model fit indices, construct validity was evaluated based on the standardized factor loading of each indicator on the latent construct. This value indicates the indicator's contribution to representing the construct being measured. In this study, an indicator was considered valid if it had a factor loading > 0.30 and was significant at $p < 0.05$ (Pergo et al., 2023; Sari et al., 2024). We further evaluated indicators with low loadings based on statistical considerations and theoretical fit before retaining or eliminating them.

We used a first-order CFA measurement model because collaborative problem-solving skills are a multidimensional construct, with several indicators that directly represent specific

dimensions without second-order factors. In this model, each item is directly associated with a single latent factor (Marsh et al., 2014).

For the four-class implementation, we summarized total and dimension-level CPS scores descriptively. Normality was examined using the Shapiro–Wilk test and homogeneity of variance using Levene’s test. Because both assumptions were supported, exploratory differences among the four intact classes were examined using one-way ANOVA followed by Tukey’s HSD for the significant omnibus result (Field, 2017). The CPS scores analyzed came from individual tests rather than shared group products. Nevertheless, students had previously participated in collaborative activities in groups of four to five and were nested within naturally occurring classes. This hierarchical structure was not explicitly modeled in the present analysis; therefore, we cannot assume independence of observations. Consequently, the ANOVA results were interpreted only as preliminary descriptive evidence of variation across implementation contexts, not as estimates of intervention effects or causal class differences. We acknowledge potential within-group and within-class dependence as a limitation that future studies should address through multilevel or cluster-adjusted analyses.

Expert comments, revision notes, implementation records, and open-ended user feedback were analyzed thematically through familiarization, initial coding, candidate-theme generation, theme review, definition and naming, and analytical reporting (Braun & Clarke, 2006). The analysis initially focused on construct representation, task and rubric clarity, system functionality, implementation feasibility, and required revisions, while remaining open to additional patterns emerging from the data. Three members of the research team independently coded the qualitative data to reduce reliance on a single researcher’s interpretation and to provide multiple perspectives on the emerging patterns. The team then compared initial codes and candidate themes, resolving disagreements through discussion and re-examination of the original data and coding definitions until consensus was reached. When needed, the team refined code definitions and reviewed affected data segments again to ensure consistent application. The resulting coding decisions and theme revisions were documented to maintain an audit trail. We addressed reflexivity

by repeatedly comparing emerging interpretations with the original comments and implementation records and by distinguishing participants’ and validators’ statements from the researchers’ interpretations. Because the qualitative strand involved a bounded set of expert comments, revision records, implementation records, and user feedback rather than iterative theoretical sampling, we did not use conventional data saturation as a stopping criterion. Instead, we assessed information adequacy by analyzing all available qualitative records relevant to the development and implementation phases and examining whether they adequately explained, complemented, or qualified the corresponding quantitative findings. The team retained representative excerpts and documented revision evidence for the major themes and reported them alongside the quantitative findings, particularly in the joint display, to provide a transparent, auditable evidentiary basis for the integrated interpretations.

The team integrated quantitative and qualitative evidence through a joint display at interpretation. The display connected Aiken’s V, expanded Gregory, ICC, CFA, class-level profiles, and implementation statistics with expert explanations, revision records, and user feedback. Evidence was interpreted as convergent, complementary, or discrepant according to whether the sources supported, extended, or qualified one another (Fetters et al., 2013).

RESULTS AND DISCUSSION

Digital Assessment System and Instrument Revisions

The developed product combined a project-based performance assessment with digital delivery, recording, and rubric-based scoring. It operationalized CPS through four related dimensions: collaboratively identifying and representing the problem, jointly planning and developing a solution, collaboratively testing and refining the solution, and communicating evidence-based results. Students addressed an authentic simple-machines problem, developed and tested a prototype, documented evidence, and communicated the resulting solution. The digital system supported task delivery, access to supporting resources, submission, scoring, and documentation of results and feedback. Figure 2 illustrates the digital CPS system’s assessment flow across these four stages.

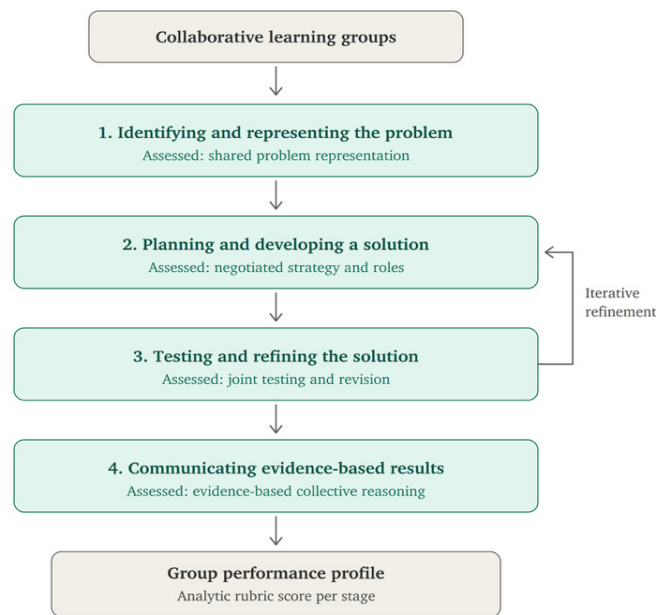


Figure 2. Example of a Digital System Display for Delivering Instruments

Figure 3 presents a simplified example of the digital system interface used to deliver the assessment instruments, specifically the Redesign module in the Sindig FMIPA learning management system for Stage 3 (testing and refining the solution). To improve readability at journal-page scale, only the main interface elements are highlighted: (1) the stage label, (2) reflective framing text, (3) task instructions, and (4) submission points for the group worksheet and process

video. The module was originally administered to participants in Indonesian. For publication, we redrew the interface at higher resolution and translated it into English; however, we preserved the instructional content, task sequence, submission requirements, and functional structure without substantive modification. The Supplementary Material provides a more detailed view of the original interface.

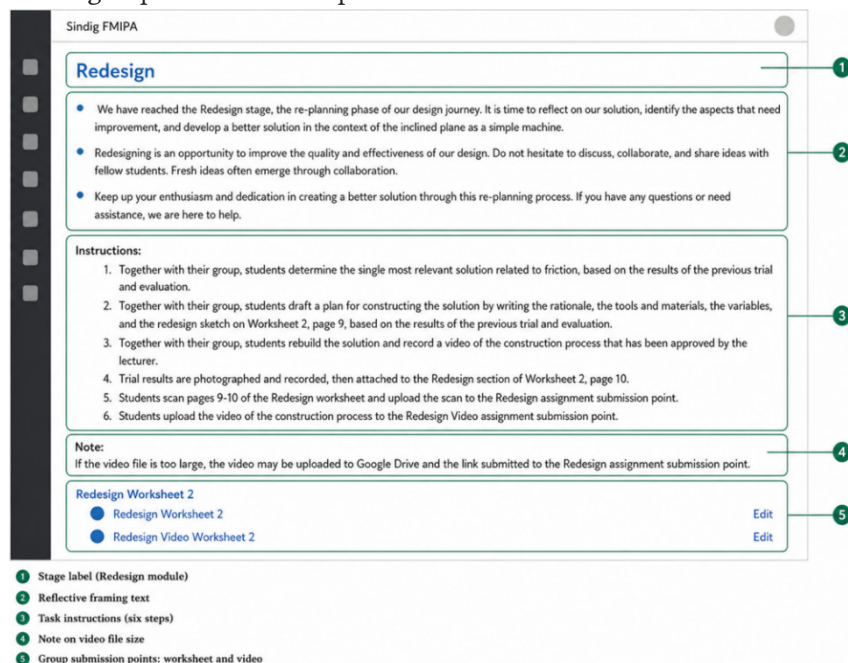


Figure 3. Example of a Digital System Interface for Delivering Instruments

Expert review identified two components requiring substantive revision. The problem-identification task initially provided insufficient conceptual exploration before groups formulated a shared problem; therefore, we added an LMS-based research-ideas activity to support conceptual synthesis and evidence-based agreement. The

initial solution-testing task ended after the first test and evaluation; we added redesign and retesting so students interpreted evidence, justified modifications, revised the prototype, and verified the revised solution. Table 1 summarizes these changes.

Table 1. Substantive Revisions to the Project-Based CPS Performance Assessment

Component	Initial form	Expert concern	Revision	Expected assessment evidence
Problem identification	Students identified an individual problem and selected one group problem.	Insufficient evidence of conceptual analysis and shared understanding.	Added LMS-based research ideas, group synthesis, and evidence-based agreement on the focal problem.	More explicit evidence of conceptual grounding, negotiation, and shared problem representation.
Solution testing	Students tested the prototype, recorded results, and evaluated the solution.	The task did not sufficiently elicit iterative evaluation and adaptation.	Added redesign, justification of modifications, revised prototype drawing, and retesting.	More explicit evidence of evaluation, strategy adaptation, and evidence-based group decision-making.

Content-Validity Evidence

The three validators' raw ratings produced corrected Aiken's V values of 0.78–1.00. Identifying and representing the problem and testing and refining the solution each obtained 0.78, whereas planning and developing a solution and communicating evidence-based results each obtained 1.00. Expanded Gregory was 1.00 for every component because all ratings were classified as strongly relevant. Thus, the components were relevant, but the lower Aiken values and expert

comments justified targeted revisions rather than a binary valid/invalid decision.

These results show that strong categorical agreement can coexist with meaningful variation in rating strength. Revision decisions therefore relied on convergence between numerical indices and expert explanations, not a single coefficient. Table 2 summarizes the content-relevance evidence and resulting revision decisions for each component.

Table 2. Content-Relevance Evidence and Resulting Decisions

CPS component	E1	E2	E3	Aiken's V	Expanded Gregory	Decision or revision
Identifying and representing the problem	3	4	3	0.78	1.00	Research-ideas activity and clearer shared-problem instructions
Planning and developing a solution	4	4	4	1.00	1.00	Retained with editorial refinement
Testing and refining the solution	4	3	3	0.78	1.00	Redesign and retesting added.
Communicating evidence-based results	4	4	4	1.00	1.00	Retained; communication rubric requires further empirical refinement.

Internal-Structure Evidence and Expert-Judgment Consistency

Whole-Sample and Exploratory Subgroup CFA

For the whole sample (N = 118), we conducted a first-order one-factor CFA using maximum likelihood (ML) estimation, with identifying, planning, testing, and communicating specified as observed indicators of the general

CPS factor. The one-factor model produced $\chi^2(2) = 5.77$, $p = 0.059$, CFI = 0.891, TLI = 0.673, GFI = 1.000, RMSEA = 0.125 (90% CI [0.000, 0.253]), and SRMR = 0.061. The nonsignificant chi-square and χ^2/df were favorable, whereas RMSEA indicated substantial residual misfit. The structure therefore received partial preliminary support rather than full confirmation. The

subgroup results were descriptive: the high-achieving group showed favorable fit, while the low-achieving group retained substantial RMSEA

misfit. Table 3 presents the complete CFA fit evidence for the whole sample and both exploratory subgroups.

Table 3. CFA Fit Evidence for the Whole Sample and Exploratory Subgroups

Analysis	n	χ^2	df	p	CFI	TLI	RM-SEA	90% CI RM-SEA	SRMR	Interpretation
Whole sample	118	5.77	2	0.059	0.891	0.673	0.125	[0.000, 0.253]	0.061	Mixed evidence; partial preliminary support
Exploratory low-achieving subgroup	63	6.02	2	0.053	0.860	0.580	0.175	[0.008, 0.351]	0.081	Mixed evidence; substantial RMSEA misfit
Exploratory high-achieving subgroup	55	1.26	2	0.533	1.000	1.282	0.000	[0.000, 0.233]	0.039	Favorable within-group fit; interpret cautiously

Because we did not test constrained multigroup models, the subgroup analyses do not demonstrate measurement invariance. We treat these contrasting subgroup patterns as exploratory findings that require replication with explicit grouping criteria and formal invariance testing.

Three dimensions exceeded the prespecified 0.30 criterion: identifying and representing the problem (I1 = 0.383, $p = 0.001$), planning

and developing a solution (I2 = 0.484, $p < 0.001$), and testing and refining the solution (I3 = 0.754, $p < 0.001$). Communicating evidence-based results was statistically significant but weaker (I4 = 0.286, $p = 0.013$). I3 was the strongest indicator, whereas I4 requires more precise performance evidence and rubric descriptors. Table 4 reports the standardized loadings for all four indicators.

Table 4. Standardized Factor Loadings

Indicator	Standardized loading	p	SE	95% CI
I1	0.383	0.001	0.111	[0.162, 0.596]
I2	0.484	0.000	0.117	[0.269, 0.728]
I3	0.754	0.000	0.144	[0.468, 1.032]
I4	0.286	0.013	0.111	[0.060, 0.496]

The standard errors (SE) ranged from 0.111 to 0.144, reflecting sampling uncertainty in the estimated standardized loadings. The 95% confidence intervals (CI) for all four indicators excluded zero, indicating that the loadings differed significantly from zero. I3 showed the highest standardized loading (0.754), although its CI was

relatively wide [0.468, 1.032], indicating uncertainty in the estimate's magnitude. I4 had the lowest loading (0.286), with a CI of [0.060, 0.496], supporting a positive but comparatively weak association with the CPS factor. Figure 4 illustrates these standardized loadings and their relative magnitudes across the four CPS dimensions.

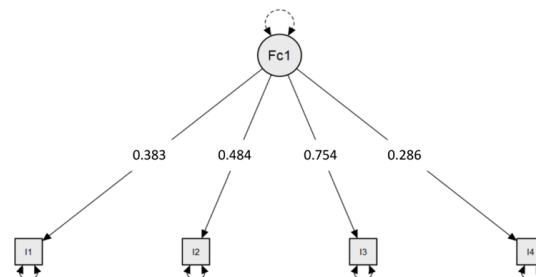


Figure 4. Standardized Loadings of the Four CPS Dimensions on the Proposed Factor

Consistency of Expert Judgments

The available expert-rating analysis yielded ICC estimates ranging from -0.103 to 0.000 . However, because the original statistical record did not retain the exact ICC model, agreement or consistency definition, measurement unit, confidence intervals, and associated significance information, these estimates cannot be interpreted as reproducible evidence of inter-validator agreement or reliability. We therefore report them only as descriptive results from the preliminary expert-validation analysis. The coefficient of variation was 11.98% , indicating relatively restricted variability among the expert ratings. The standard error of measurement (SEM) was 0.4441 , and the standard error of prediction (SEP) was 0.8131 ; we also present these values descriptively because the available statistical record does not provide sufficient model information to support formal reliability interpretation. Accordingly, the ICC, CV, SEM, and SEP were considered supplementary to the primary content-relevance evidence derived from Aiken's V , the expanded Gregory

method, and qualitative expert judgments rather than as independent evidence of measurement reliability. Importantly, these statistics concern expert ratings of the assessment components and should not be interpreted as reliability evidence for students' CPS scores. The present study did not establish inter-rater reliability or other generalizability evidence for student CPS scoring; therefore, interpret the subsequent student score profiles as preliminary.

Multiclass Implementation and User Responses

Total-Score Profiles and Class Comparison

The digital assessment system was implemented with 118 prospective science teachers in four intact classes: Class A ($n = 24$), Class B ($n = 32$), Class C ($n = 31$), and Class D ($n = 31$). No missing CPS scores were reported. Class B obtained the highest mean total score (87.0), followed by Class A (80.9), Class D (74.5), and Class C (73.3). Table 5 presents the Tukey post hoc comparisons of class total scores.

Table 5. Tukey Post Hoc Comparisons of Class Total Scores

Comparison	Mean difference	p	Decision
A–B	-6.11	0.028	Significant
A–C	7.60	0.004	Significant
A–D	6.47	0.019	Significant
B–C	13.72	< 0.001	Significant
B–D	12.58	< 0.001	Significant
C–D	-1.14	0.944	Not significant

Normality and homogeneity of variance were supported, as indicated by the Shapiro–Wilk test ($W = 0.982$, $p = 0.120$) and Levene's test, $F(3, 114) = 0.820$, $p = 0.486$. Under the conventional one-way ANOVA model, class means differed significantly, $F(3, 114) = 19.70$, $p < 0.001$, $\eta^2 = 0.341$, 95% CI $[0.199, 0.455]$. Tukey comparisons showed that Class B scored higher than Classes A, C, and D; Class A scored higher than Classes C and D; and Classes C and D did not differ significantly. However, students were nested within collaborative groups and intact classes, and this analysis did not explicitly model that dependence. Accordingly, the inferential statistics and effect-size estimate are interpreted cautiously as exploratory evidence of observed variation across naturally occurring implementation contexts rather than as evidence of treatment effects or causal class differences.

Because the four classes were naturally occurring implementation contexts rather than treatment

groups, the differences demonstrate differentiated score profiles; they do not establish intervention effectiveness or a causal class effect.

Dimension-Level Profiles

Across classes, identifying and representing the problem had the highest mean (92.77), whereas planning and developing a solution had the lowest (69.16). Testing and refinement averaged 78.73 , and communication averaged 80.04 . The pattern identifies a consistent gap between recognizing a problem and transforming shared understanding into a systematic, implementable plan. Class B showed the strongest profile in identification, planning, and communication, while Class A had the highest testing score. Classes C and D showed similar lower profiles in planning and testing. Figure 5 illustrates these dimension-level score patterns across classes.

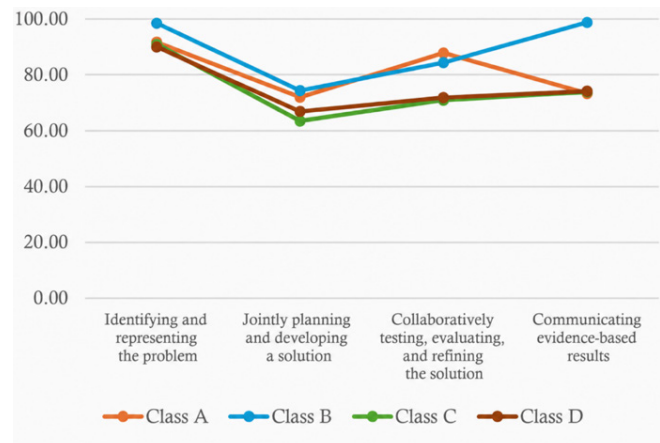


Figure 5. Mean CPS Dimension Scores Across Four Classes

User Responses and Integrated Evidence

We assessed user responses to the digital CPS assessment system using the 20-item questionnaire described in the Method section. The resulting percentages were interpreted as indicators of perceived classroom applicability and user experience, not as evidence of the validity or reliability of students' CPS performance scores.

Student responses indicated favorable perceptions across all four domains: ease of use (85.36%), assessment effectiveness (85.83%), support for collaboration and problem-solving (85.56%), and motivation and satisfaction (85.93%). The overall user-response score was

85.67%. These findings indicate favorable perceived classroom applicability and user experience with the digital CPS assessment system. However, user-response evidence reflects participants' perceptions of the system and should not be interpreted as evidence of the validity or reliability of students' CPS performance scores.

We integrated qualitative comments and revision records with the quantitative findings through a joint display. The display distinguishes areas of convergence from areas in which favorable evidence is qualified by a measurement or implementation limitation. An integration of these findings is presented in Table 6.

Table 6. Joint Display of Quantitative and Qualitative Evidence

Focus	Quantitative evidence	Qualitative explanation	Integrated inference
Problem identification	Aiken's V = 0.78; Gregory = 1.00	Experts requested stronger conceptual exploration and shared understanding.	The content was relevant, but the instructions needed strengthening to elicit the intended collaboration.
Testing and refinement	Aiken's V = 0.78; Gregory = 1.00	Experts requested redesign and retesting.	The content was relevant, but an iterative structure was needed to elicit evaluation and adaptation.
Internal structure	CFI = 0.891; TLI = 0.673; RMSEA = 0.125; SRMR = 0.061; I4 = 0.286	Communication descriptors and performance evidence were insufficiently differentiated.	The factor received partial support; communication requires further operational refinement.
Expert consistency	ICC = -0.103 to 0.000	Validators emphasized different aspects of analysis, evaluation, and revision.	Strong content relevance coexisted with inadequate expert consistency; calibration and more explicit criteria are required.
Multiclass implementation	F(3,114) = 19.7, p < 0.001	The system organized tasks, submissions, rubrics, and documentation.	The system generated differentiated profiles, but class differences are not treatment effects.
User response	Overall = 85.67%	Students viewed the system as usable and supportive.	The system was well received; usability evidence does not replace measurement evidence.

Authentic and Iterative Representation of Collaborative Problem-Solving

The system's contribution extends beyond transferring a paper-based rubric into an LMS. Empirically, expert review led to substantive revisions that strengthened the evidence elicited at key CPS stages: the problem-identification task was expanded to include conceptual exploration, group synthesis, and evidence-based agreement, while the solution-testing task was extended to include redesign, justification of modifications, and retesting. These revisions enabled the assessment workflow to document observable task evidence across shared problem representation, planning, testing, evidence-based redesign, retesting, and communication. This design is theoretically consistent with the task-dependent nature of CPS, in which different tasks can elicit different behavioral patterns and relationships with outcomes (Andrews-Todd et al., 2023; Sun et al., 2022), and with evidence that productive CPS involves problem-solving strategy, regulation, interaction, and engagement (Baucal et al., 2023). The multiclass implementation further demonstrated that the digital system could organize these assessment activities and was perceived favorably by users. However, such implementation and usability evidence should not be interpreted as evidence that the resulting CPS scores are already dependable for individual-level decisions. Because student-score inter-rater reliability or other generalizability evidence was not established, the system's capacity to generate reliable individual CPS profiles remains an intended function requiring further empirical verification. Thus, the present evidence supports the system's ability to structure and document an iterative CPS assessment process, while stronger claims concerning the reliability and interpretability of individual CPS scores should await further measurement studies.

The redesign-and-retest sequence requires groups to interpret test evidence, negotiate modifications, justify a revised design, and verify the improved solution. This design aligns with temporal and multimodal evidence that CPS emerges through changing communication, behavior, and regulation rather than a static count of participation (Ouyang et al., 2023).

Complementary Validity Evidence Rather Than a Single Valid/Invalid Decision

Aiken's *V*, expanded Gregory, CFA, and ICC answer different questions. Content indices address relevance and representativeness; CFA addresses relations among the four scores; ICC addresses consistency in expert scale use. Strong

content classification can therefore coexist with mixed structural fit and inadequate expert agreement. The corrected Aiken values also explain why the two revised components were relevant but less strongly endorsed than planning and communication. The findings support an evidence-based revision process rather than a global claim that the instrument is simply valid.

The CFA results require equal caution. The four-indicator model showed supportive and unsupportive indices, and low-degree-of-freedom models can make conventional cutoffs unstable. Fit is interpreted with loading quality, residual evidence, data characteristics, and substantive theory rather than mechanically (Goretzko et al., 2024; McNeish & Wolf, 2022). The current evidence supports continued development of the proposed structure but not a claim of fully confirmed construct validity.

Why Planning Was Most Difficult and Communication Loaded Weakly

Planning and developing a solution obtained the lowest mean score among the four CPS dimensions. However, the present qualitative and process data did not directly examine why students experienced greater difficulty at this stage. Therefore, this pattern should be treated as a post hoc theoretical interpretation rather than an empirically established mechanism. One plausible explanation is that planning places simultaneous demands on groups to integrate relevant scientific concepts, compare alternative solutions, establish design criteria, coordinate roles, and translate shared ideas into a feasible course of action. This interpretation is consistent with Baucal et al. (2023), who emphasize that productive CPS depends on problem-solving strategies and regulation directed toward both the task and group members. Accordingly, supports such as explicit design criteria, decision matrices, and evidence-based justification may strengthen planning performance, but the present study did not test their effectiveness. Future research should examine this hypothesis directly using process evidence such as interaction logs, observation records, interviews, or think-aloud data to identify the specific difficulties encountered during collaborative planning.

Communication had the weakest loading even though its mean was not the lowest. A dimension can receive relatively high scores while sharing limited variance with the general factor. The communication task may also reflect presentation fluency, digital production, speaking confidence, or visual design. Communication remains

theoretically necessary, but the rubric should distinguish solution quality, use of test evidence, scientific argumentation, organization and visualization, and each member’s collaborative contribution.

Why Strong Content Relevance Coexisted with Very Low ICC

The very low ICC does not negate content relevance. Validators may classify all components as relevant while using the four response categories differently across only four, uniformly strong objects. Restricted between-object variance can yield unstable or negative ICC estimates. Clearer anchors, calibration examples, more rateable subcomponents, independent ratings before consensus, and reporting the exact ICC model and confidence interval are therefore necessary. Student-performance reliability must also be estimated separately; expert ICC cannot substitute for consistency of student scoring.

Meaning of the Multiclass Profiles and User Responses

The significant class differences show that the system produced nonuniform and interpretable performance profiles. However, because

the classes were not randomly assigned and no intervention contrast was specified, the findings indicate between-context variation rather than effectiveness. CPS performance changes with task context (Andrews-Todd et al., 2023; Zhang & Hwang, 2023), and student teachers’ CPS composition unfolds differently across authentic online tasks (Gao et al., 2022). The present profiles should therefore support diagnosis and instructional planning rather than causal claims.

Positive user responses strengthen the applicability argument but are not validity coefficients. Their importance is that a complex performance assessment was acceptable and manageable across four classes. Integrating usability with psychometric and qualitative evidence prevents satisfaction data from masking measurement weaknesses. Joint displays are useful precisely because they make convergence, complementarity, and discrepancy explicit (Fetters & Tajima, 2022).

Novelty and Scientific Contribution

The novelty of the present study lies in its assessment architecture and implementation context rather than in a claim that digital CPS assessment is unprecedented. It is presented in Table 7.

Table 7. Comparison of the Present System with Prior Digital CPS Assessment Approaches

System/ Study	Target popu- lation	Task	Evidence source	Scoring approach	Reported validity/reli- ability evidence
PISA 2015 (OECD, 2017; Rosen et al., 2020)	15-year-old students, OECD sys- tems	Standardized human-to- agent tasks with simulat- ed partners	Selected pre- defined com- munications and actions	Individual-level, standardized psychometric framework	Large-scale standardized psychometric validation
ATC21S (Care et al., 2016; Scou- lar & Care, 2019)	School-age students, hu- man-human dyads	Human-to- human dyadic online tasks	Time-stamped actions/com- munication (log-stream data)	Behavioral in- dicators coded and modeled psychometri- cally	Psychometric modeling of process indicators
ECSPA (Hao et al., 2019; von Davier et al., 2017)	Assessment- development samples	Science-ori- ented simula- tion tasks	Collaboration process + out- come evidence, incl. text inter- action	Standardized/ computational scoring	Exploratory psychomet- ric evidence; standard- ized-assessment proto- type
Student- teacher process-data study (Gao et al., 2022)	Student teach- ers	Authentic on- line learning tasks	Process/log data	Multidimen- sional item re- sponse theory	Multidimensional struc- ture supported via pro- cess-data modeling
Present sys- tem	Prospective science teach- ers (pre-ser- vice)	Authentic sci- ence project; design-test- redesign-re- test cycle	Group artifacts + individual CPS test re- sponses	Digital analytic- rubric scoring, 4 CPS stages	Preliminary content relevance (Aiken’s V, Gregory); mixed CFA fit; expert-consistency and student-score reliability not yet established

Therefore, this work offers a transparent, practice-oriented prototype and a validity-evidence pathway for making collaboration more observable and measurable in science teacher education.

Limitations of the Current Evidence

The evidence remains preliminary and should be interpreted in light of several methodological limitations. First, content-relevance evidence was based on only three validators and four principal CPS components, making the resulting coefficients sensitive to the small number of expert judgments. Second, the one-factor CFA involved only four indicators and produced mixed fit evidence, while the exploratory subgroup analyses did not establish measurement invariance. Third, the available expert-agreement statistics did not provide a reproducible basis for a formal reliability claim, and the study did not establish inter-rater reliability or other generalizability evidence for students' CPS performance scores. This limitation matters because dependable interpretation of individual CPS profiles requires evidence that student scoring is sufficiently consistent across raters or scoring occasions. Fourth, although the analyzed CPS scores were derived from individual tests rather than directly assigning shared group-product scores to individual students, participants had previously worked within collaborative groups and were nested within four intact classes. This hierarchical dependence was not explicitly modeled; consequently, within-group or within-class correlation may have affected standard errors and between-class significance tests. Shared participation in the same collaborative activities may also have introduced common contextual influences on individual performance, even though the study did not use identical group scores as individual outcomes. Fifth, all participants were drawn from a single science-education program at one institution, which limits the generalizability of the findings to other institutional, curricular, and student populations. Sixth, the study did not examine external or criterion-related validity by relating CPS scores to independent measures of collaboration, problem solving, academic performance, or other theoretically relevant outcomes. Finally, differences among the four intact classes and favorable user responses cannot establish causal effectiveness, particularly because the classes were naturally occurring implementation contexts rather than randomly formed treatment conditions. Future research should use larger, more diverse samples; more detailed observable indicators; ca-

librated performance raters; explicit inter-rater or generalizability analyses of student CPS scores; multilevel or cluster-adjusted analyses; formal measurement-invariance testing; and external or criterion-related validity evidence. Within these limitations, the present study should be regarded as providing a transparent preliminary assessment prototype and an initial evidence pathway for making collaborative problem-solving more observable in science teacher education.

CONCLUSION

This study developed a project-based digital CPS assessment system covering four stages: problem identification and representation, solution planning and development, testing and refinement, and evidence-based communication. The system showed strong content relevance (Aiken's $V = 0.78$ – 1.00 ; expanded Gregory = 1.00), preliminary structural support with communication as the weakest dimension, significant score differences across classes, $F(3, 114) = 19.70$, $p < 0.001$, $\eta^2 = 0.341$, and favorable user responses (85.67%). Its main contribution is a practice-oriented, stage-specific digital assessment embedded in an authentic design–test–redesign–retest cycle. However, limited reliability evidence, the weak communication dimension, intact-class sampling, and unmodeled data nesting restrict stronger interpretation. Future research should refine communication descriptors, improve validator calibration and score reliability, and test structural validity and measurement invariance across larger and more diverse samples.

ACKNOWLEDGEMENTS

The authors acknowledge the Institute for Research and Community Service, Universitas Negeri Surabaya, for providing research and publication funding under grant number 454/UN38/HK/2026. The authors express deep gratitude to the Science Education Study Program, Universitas Negeri Surabaya, for the support and facilitation provided during the implementation of this research.

REFERENCES

- Adom, D., Adu-Mensah, J., & Dake, D. A. (2020). Test, measurement, and evaluation: Understanding and use of the concepts in education. *International Journal of Evaluation and Research in Education (IJERE)*, 9(1), 109–119.
- Aiken, L. R. (1985). Three Coefficients for Analyzing the Reliability and Validity of Rating. *Educational and Psychological Measurement*, 45(1),

- 131–142.
- Alenezi, M., Wardat, S., & Akour, M. (2023). The Need of Integrating Digital Education in Higher Education: Challenges and Opportunities. *Sustainability*, 15(6), 4782.
- Alfin, J., Fuad, A. Z., Nur, M., Yuanita, L., & Prahani, B. K. (2019). Development of Group Science Learning (GSL) Model to Improve the Skills of Collaborative Problem Solving, Science Process, and Self-Confidence of Primary Schools Teacher Candidates. *International Journal of Instruction*, 12(1), 147–164.
- Almanasreh, E., Moles, R. J., & Chen, T. F. (2022). A practical approach to the assessment and quantification of content validity. In *Contemporary Research Methods in Pharmacy and Health Services* (pp. 583–599). Elsevier.
- Anchunda, H. Y., & Kaewurai, W. (2025). An instructional model development based on inquiry-based and problem-based approaches to enhance prospective teachers' teamwork and collaborative problem-solving competence. *Social Sciences & Humanities Open*, 11, 101480.
- Andrews-Todd, J., Jiang, Y., Steinberg, J., Pugh, S. L., & D'Mello, S. K. (2023). Investigating collaborative problem solving skills and outcomes across computer-based tasks. *Computers & Education*, 207, 104928.
- Baucal, A., Jošić, S., Stepanović Ilić, I., Videnović, M., Ivanović, J., & Krstić, K. (2023). What makes peer collaborative problem solving productive or unproductive: A qualitative systematic review. *Educational Research Review*, 41, 100567.
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101.
- Brod, M., Pohlman, B., & Waldman, L. T. (2023). Qualitative Research and Content Validity. In *Encyclopedia of Quality of Life and Well-Being Research* (pp. 5658–5666). Springer International Publishing.
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford publications.
- Brown, T. A., & Moore, M. T. (2012). Confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 361–379). Guilford Press.
- Care, E., Scoular, C., & Griffin, P. (2016). Assessment of Collaborative Problem Solving in Education Environments. *Applied Measurement in Education*, 29(4), 250–264.
- Carrillo-Avalos, B. A., Leenen, I., Trejo-Mejía, J. A., & Sánchez-Mendiola, M. (2025). Bridging Validity Frameworks in Assessment: Beyond Traditional Approaches in Health Professions Education. *Teaching and Learning in Medicine*, 37(2), 229–238.
- Chai, H., Hu, T., & Wu, L. (2024). Computer-based assessment of collaborative problem solving skills: A systematic review of empirical research. *Educational Research Review*, 43, 100591.
- Chen, L., Inoue, K., Goda, Y., Okubo, F., Taniguchi, Y., Oi, M., Konomi, S., Ogata, H., & Yamada, M. (2020). Exploring Factors that Influence Collaborative Problem Solving Awareness in Science Education. *Technology, Knowledge and Learning*, 25(2), 337–366.
- Collier, J. E. (2026). *Applied Structural Equation Modeling Using AMOS: Basic to Advanced Techniques* (2nd Ed). Routledge.
- Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.
- Fauziah, A. N. M., Wasis, Hendratmoko, A. F., Mahdiannur, M. A., Ermawan, M. Z. F., Suwandi, E., & Ratri, S. Y. (2024). Relationship Between Critical Thinking and Scientific Argumentation in Science Learning. *Jurnal Pendidikan IPA Indonesia*, 13(2), 239–254.
- Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving Integration in Mixed Methods Designs—Principles and Practices. *Health Services Research*, 48(6pt2), 2134–2156.
- Fetters, M. D., & Tajima, C. (2022). Joint Displays of Integrated Data Collection in Mixed Methods Research. *International Journal of Qualitative Methods*, 21.
- Field, A. (2017). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE.
- Gao, Q., Zhang, S., Cai, Z., Liu, K., Hui, N., & Tong, M. (2022). Understanding student teachers' collaborative problem solving competency: Insights from process data and multidimensional item response theory. *Thinking Skills and Creativity*, 45, 101097.
- García-Carmona, A. (2020). From Inquiry-Based Science Education to the Approach Based on Scientific Practices. *Science & Education*, 29(2), 443–463.
- Gillies, R. M. (2023). Teaching Science That Is Inquiry-Based: Practices and Principles. In *Challenges in Science Education* (pp. 39–58). Springer International Publishing.
- Goretzko, D., Siemund, K., & Sterner, P. (2024). Evaluating Model Fit of Measurement Models in Confirmatory Factor Analysis. *Educational and Psychological Measurement*, 84(1), 123–144.
- Hao, J., Liu, L., Kyllonen, P., Flor, M., & von Davier, A. A. (2019). Psychometric Considerations and a General Scoring Strategy for Assessments of Collaborative Problem Solving. *ETS Research Report Series*, 2019(1), 1–17.
- Hariyono, E., Rosdiana, L., Budiyanto, M., Satriawan, M., Nabilah, H., & Rizki, I. A. (2026). Factors Determining Socio-Scientific Issues in STEAM Education to Enhance Problem-Solving Skills for Pre-Service Teachers: Development and Validation of a Measurement Model. *Jurnal Pendidikan IPA Indonesia*, 15(1), 14–28.
- Haryani, E., Coben, W. W., Pleasants, B. A.-S., & Fetters, M. K. (2021). Analysis of Teachers' Resources for Integrating the Skills of Creativity and Innovation, Critical Thinking and Problem Solving, Collaboration, and Communication in

- Science Classrooms. *Jurnal Pendidikan IPA Indonesia*, 10(1), 92–102.
- Hendratmoko, A. F., Fauziah, A. N. M., Susiyawati, E., Astriani, D., Nurita, T., Martini, Erman, Septaria, K., Fi'aunillah, T. K., & Boey, A. (2026). Integrating ChatGPT-driven self-assessment and Debate Based on Inquiry Learning (DBOIL) to enhance students' argumentation skills: Leveraging AI for quality science learning aligned with SDG 4. *E3S Web of Conferences*, 696, 02005.
- Hendratmoko, A. F., Madlazim, M., Widodo, W., & Astutik, S. (2024). Debate Based on Inquiry Learning (DBOIL): An Innovative Learning Model to Improve Students' Skills in Scientific Argumentation. *TEM Journal*, 13(4), 3295–3303.
- Hidayat, R., Habibi, A., Mohd Saad, M. R., Mukminin, A., & Wan Idris, W. I. bin. (2019). Exploratory and confirmatory factor analysis of PERMA for Indonesian students in mathematics education programmes. *Pedagogika*, 132(4), 147–165.
- Kaspi, S., & Venkatraman, S. (2023). Data-Driven Decision-Making (DDDM) for Higher Education Assessments: A Case Study. *Systems*, 11(6), 306.
- Kline, R. B. (2023). *Principles and practice of structural equation modeling*. Guilford publications.
- Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2), 155–163.
- Lin, L., Zhou, D., Wang, J., & Wang, Y. (2024). A Systematic Review of Big Data Driven Education Evaluation. *Sage Open*, 14(2).
- MacCallum, R. C., & Austin, J. T. (2000). Applications of Structural Equation Modeling in Psychological Research. *Annual Review of Psychology*, 51(1), 201–226. U
- Magaji, A. (2021). Promoting Problem-Solving Skills among Secondary Science Students through Problem Based Learning. *International Journal of Instruction*, 14(4), 549–566.
- Maier, U., & Klotz, C. (2022). Personalized feedback in digital learning environments: Classification framework and literature review. *Computers and Education: Artificial Intelligence*, 3, 100080.
- Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory Structural Equation Modeling: An Integration of the Best Features of Exploratory and Confirmatory Factor Analysis. *Annual Review of Clinical Psychology*, 10(1), 85–110.
- McNeish, D., & Wolf, M. G. (2022). Dynamic fit index cutoffs for one-factor models. *Behavior Research Methods*, 55(3), 1157–1174.
- Meijer, H., Hoekstra, R., Brouwer, J., & Strijbos, J.-W. (2020). Unfolding collaborative learning assessment literacy: a reflection on current assessment methods in higher education. *Assessment & Evaluation in Higher Education*, 45(8), 1222–1240.
- Mutmainna, M., Hasyim, M., Syamsuriwal, S., Setiaji, B., & Setiawati, F. A. (2024). Modeling of Colorado Learning Attitude Science Survey in Indonesian Version: A Study with Applying Item Response Theory. *JP3I (Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia)*, 13(1), 80–104.
- Negrón, J., López-Cruz, M., Felmer, P., & Sepúlveda, C. L. (2026). Collaborative problem solving in science classrooms: Tools for developing 21st century skills. *School Science and Mathematics*, 126(3), 235–250.
- Nugraha, M. F. (2022). The Development and Validation of Jihad Academic Scale (JAS). *JP3I (Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia)*, 11(2), 182–195.
- Nurita, T., Fauziah, A. N. M., Mahdiannur, M. A., Susilowati, R., & Hendratmoko, A. F. (2025). Fostering Creative Thinking Skills in Creating Domain of Anderson's Taxonomy for Prospective Science Teachers Using E-Modul. *TEM Journal*, 14(3), 2339–2348.
- OECD. (2017). *PISA 2015 Results (Volume V): Collaborative Problem Solving*. OECD Publishing.
- Ouyang, F., Dinh, T. A., & Xu, W. (2023). A Systematic Review of AI-Driven Educational Assessment in STEM Education. *Journal for STEM Education Research*, 6(3), 408–426.
- Pergo, A. T. C., França, A. B., Teixeira, P. H. M., Siqueira, L. B., Cortez, P. J. O., & Vitorino, L. M. (2023). Development and validation of a vaccine hesitancy scale for COVID-19 to the Portuguese language. *Research, Society and Development*, 12(3), e4012340357.
- Retnawati, H. (2016). Proving content validity of self-regulated learning scale (The comparison of Aiken index and expanded Gregory index). *REID (Research and Evaluation in Education)*, 2(2), 155–164.
- Richey, R. C., & Klein, J. D. (2007). *Design and development research: Methods, strategies, and issues*. Lawrence Erlbaum Associates.
- Rosen, Y., Wolf, I., & Stoeffler, K. (2020). Fostering collaborative problem solving skills in science: The Animalia project. *Computers in Human Behavior*, 104, 105922.
- Sajidan, S., Suranto, S., Atmojo, I. R. W., Saputri, D. Y., & Etviana, R. (2022). Problem-Based Learning-Collaboration (PBL-C) Model in Elementary School Science Learning in the Industrial Revolution Era 4.0 and Indonesia Society 5.0. *Jurnal Pendidikan IPA Indonesia*, 11(3), 477–488.
- Sari, C. M., Ridwan, A. R., & Sastrawijaya, Y. S. (2024). Learning Environment of Numeracy: Development and Validation With Confirmatory Factor Analysis and Rasch Model. *Pedagogika*, 153(1), 185–198.
- Scoular, C., & Care, E. (2019). A Generalized Scoring Process to Measure Collaborative Problem Solving in Online Environments. *Educational Assessment*, 24(3), 213–234.

- Sincovich, A., Gregory, T., Zanon, C., Santos, D. D., Lynch, J., & Brinkman, S. A. (2019). Measuring early childhood development in multiple contexts: the internal factor structure and reliability of the early Human Capability Index in seven low and middle income countries. *BMC Pediatrics*, 19(1), 471.
- Siswanto, J., Khoiri, N., & Saphira, H. V. (2025). The Effectiveness of the Algodoo-Assisted IBMRO Model in Improving Physics Problem-Solving Skills. *Jurnal Pendidikan IPA Indonesia*, 14(3), 496–508.
- Sun, C., Shute, V. J., Stewart, A. E. B., Beck-White, Q., Reinhardt, C. R., Zhou, G., Duran, N., & D'Mello, S. K. (2022). The relationship between collaborative problem solving behaviors and solution outcomes in a game-based learning environment. *Computers in Human Behavior*, 128, 107120.
- Sun, C., Shute, V. J., Stewart, A., Yonehiro, J., Duran, N., & D'Mello, S. (2020). Towards a generalized competency model of collaborative problem solving. *Computers & Education*, 143, 103672.
- Tang, T., Vezzani, V., & Eriksson, V. (2020). Developing critical thinking, collective creativity skills and problem solving through playful design jams. *Thinking Skills and Creativity*, 37, 100696.
- Thornhill-Miller, B., Camarda, A., Mercier, M., Burkhardt, J.-M., Morisseau, T., Bourgeois-Bougrine, S., Vinchon, F., El Hayek, S., Augereau-Landais, M., Mourey, F., Feybesse, C., Sundquist, D., & Lubart, T. (2023). Creativity, Critical Thinking, Communication, and Collaboration: Assessment, Certification, and Promotion of 21st Century Skills for the Future of Work and Education. *Journal of Intelligence*, 11(3), 54.
- Tomar, S., Arundhati, A., Gupta, S., & Sharma, M. (2024). Digital assessment: impact on student motivation, peer learning, group dynamics. *Journal of Education and Learning (EduLearn)*, 18(1), 9–17.
- von Davier, A. A., Hao, J., Liu, L., & Kyllonen, P. (2017). Interdisciplinary research agenda in support of assessment of collaborative problem solving: lessons learned from developing a Collaborative Science Assessment Prototype. *Computers in Human Behavior*, 76, 631–640.
- Widodo, W., Rosdiana, L., Sari, D. A. P., Sari, D. P., Purnomo, A. R., Roqobih, F. D., & Hendratmoko, A. F. (2026). Activity-Based-Lesson Learn-Reflection (ALLR) Assisted by HTML5 Package (H5P) to Develop Critical Thinking Skills of Future Science Teachers. *TEM Journal*, 15(2), 1826–1841.
- Wilson, M. (2023). *Constructing Measures: An Item Response Modeling Approach* (2nd ed.). Routledge.
- Zhang, D., & Hwang, G.-J. (2023). Effects of Interaction between Peer Assessment and Problem-Solving Tendencies on Students' Learning Achievements and Collaboration in Mobile Technology-Supported Project-Based Learning. *Journal of Educational Computing Research*, 61(1), 208–234.