PROBLEM-BASED LEARNING AND CRITICAL THINKING IN ISLAMIC CULTURAL HISTORY

Riza Natasya, Umi Kalsum, Parlaungan Lubis
Universitas Islam Sumatera Utara, Sumatera Utara, Indonesia
Email. rizanatasya55@gmail.com, umikalsum@fai.uisu.ac.id, parlaungan.lubis@fai.uisu.ac.id

Abstrak

Penelitian ini menganalisis penerapan Problem-Based Learning (PBL) dan perbedaan hasil pengukuran berpikir kritis pada pembelajaran Sejarah Kebudayaan Islam di MTs Al-Ulum Medan. Desain kuasi-eksperimen kelompok kontrol nonekuivalen melibatkan 51 siswa kelas VIII, terdiri atas 26 siswa kelas eksperimen dan 25 siswa kelas kontrol. Data berasal dari tes esai, angket, lembar observasi, serta dokumentasi. Rerata tes esai meningkat dari 52,69 menjadi 75,73 pada kelompok PBL dan dari 46,72 menjadi 54,24 pada kelompok kontrol; kedua kelompok sudah berbeda 5,97 poin sebelum pembelajaran. N-Gain yang dilaporkan masing-masing sebesar 0,3671 dan 0,2705. Secara terpisah, keluaran uji-t menunjukkan t(49) = 2,503 dan p = 0,016 dengan rerata posttest 27,54 dan 24,56, bukan skala tes esai yang dilaporkan. Oleh sebab itu, signifikansi tersebut belum dapat dikaitkan secara pasti dengan selisih skor esai. Temuan menunjukkan keuntungan deskriptif pada kelas PBL, tetapi keterbatasan kesetaraan awal, dokumentasi pelaksanaan, dan keterlacakan instrumen menghalangi kesimpulan kausal. Verifikasi data mentah dan rubrik penilaian diperlukan sebelum menetapkan besaran pengaruh.

Kata kunci: Pembelajaran Berbasis Masalah, Berpikir Kritis, Kuasi-Eksperimen

Abstract

This study examines problem-based learning (PBL) and differences in measured critical-thinking performance in Islamic Cultural History at MTs Al-Ulum Medan. A nonequivalent control-group quasi-experiment involved 51 eighth-grade students, with 26 in the PBL class and 25 in the comparison class. Evidence comprised essay tests, questionnaires, observation sheets, and documentation. Reported essay-test means increased from 52.69 to 75.73 in the PBL group and from 46.72 to 54.24 in the control group; the groups already differed by 5.97 points at baseline. Reported normalized gains were 0.3671 and 0.2705, respectively. A separate independent-samples test yielded t(49) = 2.503, p = .016, but its posttest means of 27.54 and 24.56 do not match the reported essay-test scale. Consequently, the significant test cannot yet be unequivocally attributed to the essay-score comparison. The descriptive pattern favors the PBL class, while initial nonequivalence, limited implementation records, and unreconciled measurement scales preclude a firm causal conclusion. Original score files, instrument definitions, and scoring rubrics need verification.

Keywords: Problem-Based Learning, Critical Thinking, Quasi-Experiment

INTRODUCTION

Critical thinking is central to education that expects learners to evaluate claims, examine evidence, compare explanations, and justify conclusions rather than merely reproduce information. Ennis (1991) characterizes critical thinking as reasonable and reflective judgment about what to believe or do, while Kuhn (1999) emphasizes the developmental importance of evaluating knowledge claims and reasoning strategies. In secondary classrooms, these capacities are consequential because learners increasingly encounter competing narratives in textbooks, online materials, and everyday discussions. Instruction that cultivates inquiry and reasoned argument can support both disciplinary understanding and students' ability to use information responsibly. Meta-analytic evidence suggests that critical thinking benefits from explicit instructional attention and opportunities for dialogue, although the strength of effects differs across interventions (Abrami et al., 2015).

Islamic Cultural History, known in Indonesian madrasahs as Sejarah Kebudayaan Islam (SKI), offers substantial material for such inquiry. Historical narratives encompass changes in institutions, encounters among communities, political choices, intellectual movements, and interpretations of social transformation. Students therefore need more than chronological recall: they can examine the circumstances surrounding decisions, compare historical perspectives, distinguish documented claims from assumptions, and consider the implications of past events. These activities remain consistent with the subject's religious and educational orientation without treating historical interpretation as a predetermined answer. Yet a lecture-dominated approach risks reducing complex historical material to names, dates, and summaries that students are expected to memorize.

The local problem requires a more precise distinction between low general achievement and weak critical reasoning. Preliminary observations at MTs Al-Ulum Medan reported an average SKI score of 68 against a mastery criterion of 75, with 62% below that criterion. The source also estimated that approximately 30% of students contributed questions or comments and that classroom time was dominated by lectures (85%) rather than active learning (15%). These figures describe the study's reported context; neither the observation period, the denominator underlying these percentages, nor their direct correspondence to the two research classes can be independently established from the supplied account. Low subject scores and infrequent questioning are reasons to investigate instructional practice, not direct measurements of students' ability to evaluate historical evidence or justify conclusions. The research problem is therefore whether PBL provides opportunities for more explicit historical reasoning, and whether the available instruments can detect that reasoning reliably.

Problem-based learning (PBL) offers a possible instructional response by organizing learning around a problem that requires investigation, discussion, formulation of explanations, and reflection. Savery (2006) distinguishes genuine PBL from simply assigning a problem after a lecture: learners are expected to identify what they need to know and to construct defensible solutions. Hmelo-Silver (2004) explains that facilitated problem solving can develop flexible knowledge and self-directed learning, while Schmidt et al. (2011) connect PBL processes with the activation and elaboration of prior knowledge. Students encounter uncertainty rather than only prepared answers, and the teacher supports progress through questions, resources, and feedback.

In SKI, a PBL task should make a historical explanation contestable on evidential grounds. For example, learners might compare plausible explanations for a political transition, identify what each source can support, and defend a provisional conclusion while acknowledging missing

information; this is a proposed illustration, not a documented task from the intervention. The five Ennis-related domains used in the thesis can be aligned with such work: elementary clarification identifies the historical question; basic support assesses sources and reasons; inference draws a conclusion; further clarification exposes assumptions; and strategies and tactics guide investigation or revision (Ennis, 1991, 2015). This mapping specifies what an assessment rubric would need to capture, but the article does not contain the complete question-by-indicator blueprint, scoring descriptors, or rated student responses. Thus, a rise in aggregated essay scores cannot by itself reveal which dimension of critical thinking improved or whether the rubric rewarded historical justification rather than factual recall.

The literature offers no reason to assume that PBL has uniform effects across subject matter and assessment design. Positive syntheses (Dochy et al., 2003; Gijbels et al., 2005; Walker & Leary, 2009) coexist with more qualified or inconsistent results (Manuaba et al., 2022; Vecolitha & Yudiarso, 2021). The present study addresses a narrower question: how a five-phase PBL approach was reported in a Grade VIII SKI context and how two intact classes compared on available outcome records. Its contribution is the educationally specific use of historical reasoning indicators, together with transparent recognition of an important measurement problem: the essay-based descriptive means and the independent-samples SPSS output are on different reported scales. The primary intended construct is critical-thinking performance in SKI; questionnaire responses are supporting evidence and must not be treated as a substitute. Before any confirmatory hypothesis about treatment effects is asserted, the authors must identify the primary score variable and reconcile its test results with the raw data.

Instructional support is particularly important when students are unfamiliar with inquiry-oriented tasks. Hmelo-Silver et al. (2007) argue that PBL commonly involves scaffolding rather than unguided discovery. Rotgans and Schmidt (2011) show how cognitive engagement can fluctuate during problem-centered learning. Such observations imply that a local evaluation should examine not only test-score differences but also the way PBL is implemented and assessed. A promising approach may appear ineffective if learners do not understand the problem or lack sufficient supporting resources, while apparent gains may be distorted if outcome measures do not correspond closely to critical-thinking processes.

This study examines the classroom design and reported outcome differences between PBL and conventional teaching of Islamic Cultural History in two intact Grade VIII classes. It documents a possible route from problem-oriented historical inquiry to evidence-informed explanation, while testing the limits of available statistical claims. The research questions concern the stated implementation sequence, changes in essay scores across the two classes, and the interpretive consequences of nonequivalent starting levels and multiple assessment scales. No wholly new model of PBL is claimed. Rather, the local contribution is to specify what can be learned from the reported comparisons and what further documentation is necessary to establish stronger inferences (Savery, 2006; Hmelo-Silver, 2004; Sugiyono, 2019).

METHOD

The research adopted a quantitative nonequivalent control-group pretest–posttest design. Grade VIII-2 (n = 26) received PBL and Grade VIII-3 (n = 25) received conventional SKI instruction; all 51 students were included in the reported before-and-after summaries. Group allocation followed existing classes, not individual random assignment. The source does not document objective criteria for selecting these particular classes or demonstrate that teacher,

timetable, prior knowledge, and learning resources were equivalent. Baseline essay means were 52.69 and 46.72, a 5.97-point difference. This disparity should be addressed in any eventual treatment comparison, for example through an appropriately justified adjusted posttest analysis if the underlying observations and assumptions permit. Since no such reanalysis is supported by the available student-level data, this article reports descriptive comparisons and the source's original inferential output separately.

The planned PBL sequence comprised five phases: orientation to a historical problem, organization for learning, guided investigation, preparation and presentation of explanations, and reflection on problem solving. The comparison class received conventional teaching. The original source does not provide a sufficiently detailed meeting-by-meeting record, the precise historical problems used, observer agreement, or a fidelity checklist showing how often each phase took place. Accordingly, these steps describe the stated treatment procedure rather than independently validated implementation. For a reproducible report, the authors should provide the lesson schedule, topic, teaching personnel for both groups, examples of the original problem prompts, and completed observation records; these details should be supplied only from authentic field documentation. Without them, differences cannot safely be attributed to particular PBL mechanisms.

The study reports essay-based pretests and posttests, student-response questionnaires, observation sheets, and documentary records. Critical thinking was defined through Ennis-inspired indicators of clarification, basic support, inference, further explanation, and strategic reasoning. Two distinct ten-item instrument summaries show Cronbach's alpha values of .682 and .700 and reported item validity checks, but the supplied tables do not resolve unequivocally which scale generated each outcome. Internal consistency alone does not establish the construct validity of an essay measure. A defensible account requires the item-to-indicator matrix, maximum scores, essay scoring rubric, scoring procedure (including any second rater), and explicit identification of which instrument produced the descriptive essay means, normalized gains, and SPSS comparison. These details cannot be reconstructed from aggregate results and have not been invented here.

The source reports distributional tests, Levene's variance check, an independent-samples t-test, a paired questionnaire test, and normalized gain. The independent-samples output gives $t(49) = 2.503$, $p = .016$, comparing means of 27.54 and 24.56 (posttest Levene $p = .754$). The paired questionnaire comparison gives $t(25) = -2.641$, $p = .014$, for means of 27.92 and 29.65 when calculated in pre-minus-post order. These statistics should be attached only to the variables actually entered into the tests; they cannot automatically establish significance for the separate essay summary of 75.73 versus 54.24. The source reports N-Gain values of .3671 and .2705, but the paired scores and score maxima needed for independent verification are not available in the article. No ANCOVA, confidence interval, adjusted difference, or cross-scale effect size is calculated without the underlying observations. The appropriate publication requirement is reconciliation of instrument labels, units, original SPSS variable definitions, and anonymized raw scores.

RESULTS AND DISCUSSIONS

Implementation of PBL and the Assessment of Historical Reasoning

The intended five-phase PBL arrangement asked Grade VIII SKI students to identify a historical problem, gather and interpret information, present an explanation, and review their reasoning. This can create useful opportunities to distinguish assertions from evidence, but the

presence of group work is not proof of critical thinking. The original account does not reproduce an actual historical problem assigned during the intervention or provide graded samples of learners' responses. Thus, the stated PBL sequence should be recognized as the instructional design, not equated with independently observed gains in argument quality. The relation to critical-thinking outcomes requires evidence that students explicitly compared historical explanations, evaluated source credibility, and justified a reasoned conclusion (Savery, 2006; Hmelo-Silver, 2004).

The research setting made this intervention meaningful because preliminary observations indicated limited active questioning and reliance on teacher talk. A problem-oriented lesson has the potential to redistribute some intellectual responsibility toward learners, yet its success depends on the quality of questions, the availability of understandable sources, and how teachers handle uncertainty. Savery (2006) treats student inquiry as a defining feature of PBL, whereas Hmelo-Silver (2004) stresses facilitation and reflection. In practice, students new to analytical discussion may need prompts to distinguish evidence from an opinion, formulate historical claims, and explain how the available information supports an inference. Without such support, a problem can produce lively conversation without necessarily developing critical reasoning.

A clearer operational framework links PBL phases to the five critical-thinking domains in the research blueprint. Problem orientation can elicit elementary clarification; investigation can require evaluation of basic support; presentation can expose inferences and assumptions; and reflection can involve judging strategies and revising conclusions (Ennis, 1991, 2015). This alignment is analytic and should not be mistaken for measured performance of individual students. Neither the item-by-domain essay mapping nor a rubric showing how evidence quality and argument coherence were scored accompanies the published narrative. Furthermore, the ten-item reliability summaries cannot establish the quality of the essay task unless the underlying instrument type is identified. The results should avoid describing critical thinking as a single fully verified outcome before these links are documented.

Instrument documentation remains a central weakness rather than a minor technicality. The source lists two ten-item sets and alpha values of .682 and .700, but does not clearly identify which belongs to the critical-thinking essay assessment and which to the student-response questionnaire. Consequently, the observed variation could reflect distinct constructs. To permit independent examination, the authors should append the anonymized item blueprint, score range, grading descriptors, original validity and reliability outputs, and a variable dictionary mapping every table to its instrument. In the absence of these records, reliability coefficients are described as reported, not treated as verification of critical-thinking measurement. Student questionnaires may inform perceived engagement but cannot replace scored historical argumentation.

The need for explicit guidance is consistent with Hmelo-Silver et al. (2007), who distinguish scaffolded inquiry from minimally guided discovery. Schmidt et al. (2011) further explain how activation of prior knowledge and elaboration through discussion are central to productive PBL. Rotgans and Schmidt (2011) identify engagement as a process shaped by classroom interactions rather than a stable attribute created automatically by the teaching model. Thus, the present implementation should be read as a plausible instructional pathway: PBL gave students opportunities to work with complex tasks, but the source does not quantify the quality or frequency of the observed reasoning moves.

The case offers a promising disciplinary rationale for PBL: SKI inquiry can be built around conflicting interpretations of historical events rather than memorizing a single chronology. Still, instructional feasibility, fidelity, and demonstrated thinking are distinct outcomes. The study does

not provide a count of speaking turns, observer ratings across PBL phases, or comparable classroom artifacts for experimental and control groups. Interpretation should therefore remain at the level of a documented instructional plan with reported classroom activities. Future classroom research can strengthen the evidence through actual lesson questions, observation sheets, anonymized pupil work, and explicit scoring criteria. Those materials should be recovered from records, not retrospectively reconstructed as if observed.

Pretest–Posttest Differences Between Experimental and Control Classes

The thesis reports an essay-test mean of 52.69 at pretest and 75.73 at posttest for 26 PBL students, a descriptive rise of 23.04 points. For the 25 comparison students, the corresponding means were 46.72 and 54.24, a rise of 7.52 points. Thus, the observed mean increase was 15.52 points larger in the PBL class. These are arithmetic differences of reported group means, not an adjusted causal effect. The experimental class entered with a 5.97-point advantage, and the intact-class design offers no assurance that all other influences were balanced. Student-level change scores or a properly specified baseline-adjusted analysis would be needed before determining whether the difference can be assigned to the PBL approach.

Because the classes were selected intact rather than randomly constituted, their initial difference may reflect prior SKI knowledge, learning habits, or other classroom circumstances. A statistically nonsignificant baseline difference, even if one were established, would not prove substantive equivalence. The original article lacks raw student scores and a verified adjusted comparison. Descriptive pretest–posttest gains are therefore informative about the participating classes, but do not establish the magnitude of the treatment effect. When original data are available, the author should identify a single primary outcome, investigate baseline imbalance and model assumptions, and consider appropriate adjusted group comparison. No adjusted estimate is asserted here. The broader PBL literature likewise shows variations across implementations and measurements (Gijbels et al., 2005; Walker & Leary, 2009).

The most consequential reporting discrepancy is between two posttest score scales. The essay summary gives 75.73 for PBL and 54.24 for control, whereas a separate SPSS table lists 27.54 (SD = 4.528, n = 26) and 24.56 (SD = 3.938, n = 25). The SPSS table reports a mean difference of approximately 2.978, $t(49) = 2.503$, $p = .016$, with Levene $p = .754$. These reported statistics are compatible with one another on that particular scale; they cannot be taken as a significance test of the 21.49-point essay posttest difference unless a verified transformation or instrument link is provided. The author should recover the original SPSS output, variable names, score coding, and raw records. Until then, the significant SPSS result and larger descriptive essay gain must be reported as separate findings.

The source reports Shapiro–Wilk p values of .209 and .126 at pretest and .129 and .082 at posttest, all exceeding .05. However, the assumption checks support a particular analysis only when linked to the same outcome variable and group records. A distribution test does not establish that essay scores and the SPSS outcome represent the same construct or scale. Likewise, variance homogeneity does not eliminate selection bias arising from nonequivalent classes. These limitations are explicitly retained here to prevent the inferential result from appearing stronger than the documentation warrants.

The results are educationally plausible in light of the task characteristics of PBL. Students who investigate a historical problem might need to organize information and defend an interpretation, activities that may support more successful written responses. Hmelo-Silver (2004)

and Abrami et al. (2015) provide a theoretical and empirical rationale for linking guided inquiry with reasoning development. Nevertheless, broader reviews caution that the advantage of PBL is neither uniform nor independent of measurement. Strobel and van Barneveld (2009) found different patterns depending on whether assessments emphasize retention or skills, while Manuaba et al. (2022) reported no significant advantage in a different educational population. Direct comparison with these studies is limited by differences in context and instruments.

On the evidence available, two narrowly stated findings can be maintained: the essay-test summaries show a larger descriptive gain in the PBL class, and the separately reported SPSS comparison detects a difference on a score scale not reconciled with those summaries. The article cannot yet establish that $t(49) = 2.503$ is the inferential test of the essay outcome or that 23.04 points represent a causally identified PBL gain. Before final publication, the author must specify the instrument behind each table, provide anonymized paired scores and rubrics, and repeat the intended main comparison on the correct dataset. This is a condition for strengthening inference, not a new calculation supplied by the reviewer.

Normalized Gain, Student Responses, and Implications for SKI Teaching

The original report gives mean normalized gains of .3671 for the PBL class and .2705 for the control class. Under the categories adopted in the source, these are moderate and low, respectively. N-Gain is sensitive to the specified score maximum, students' starting values, and whether individual gains are averaged or gain is computed from class averages. Because a complete anonymized pretest–posttest dataset and the exact operational scoring convention are not available in this journal article, these values remain reported statistics, not independently validated effect sizes. Moreover, a difference in N-Gain does not resolve the nonequivalent baseline or the separate SPSS-scale discrepancy. It is inappropriate to equate the gain difference with a causal percentage advantage of PBL.

The initial essay scores differ by 5.97 points, making the unequal starting position relevant to both raw differences and normalized gain. Further, group means do not permit confirmation of the mean of each student's normalized gain. Any future revised statistical table should identify the maximum possible essay score, the paired student scores used, the formula and aggregation rule, the resulting uncertainty, and whether all statistics refer to the same instrument. If these records cannot be supplied, it is methodologically preferable to retain the reported gains with a caution rather than substitute an unverified recalculation. The same rule applies to claims of large treatment effects.

A separate student-response questionnaire is reported for the 26 PBL students, with means rising from 27.92 to 29.65. The source gives a paired-samples result of $t(25) = -2.641$, $p = .014$ for the pre-minus-post comparison. This indicates a reported difference in that measured scale, but the scale's construct and item coding must be identified before interpreting it as engagement, acceptance, or critical thinking. In particular, more positive responses to PBL are not automatically equivalent to stronger ability to evaluate historical evidence. The article therefore treats questionnaire scores as supplementary and does not combine them with essay-test gains or use the paired result as a test of the intervention's effect against the control group.

The distinction between engagement and higher-order performance is supported by studies of PBL processes. Rotgans and Schmidt (2011) observed that cognitive engagement can change within a problem-based learning experience, and Schmidt et al. (2011) highlighted the combined contributions of cognitive processing and interest. The educational implication is not

simply that group discussion is desirable, but that tasks should demand evidence-informed interpretation. In SKI lessons, teachers could ask students to examine the causes of a historical development, contrast accounts of an event, and explain the assumptions underlying their preferred conclusion. Assessment criteria should explicitly reward justified inference instead of only rewarding correct factual recall. Rotgans et al. (2011) describe a related line of work on the learning process in concentrated problem-based formats.

External syntheses help situate the result without overstating its generality. Harahap et al. (2023) examined critical-thinking outcomes in secondary biology and found a basis for exploring PBL as a learning approach; Aba-Oli et al. (2025) similarly addressed higher-order thinking in secondary mathematics. Shofa et al. (2026) explored how educational setting and implementation may moderate results. Conversely, Vecolitha and Yudiarso (2021) and Manuaba et al. (2022) report evidence that challenges any claim of automatic superiority. These mixed findings reinforce the need for local curricular adaptation and for clear measures rather than an assumption that all PBL variants yield the same gains.

For the participating madrasah, the immediate practical implication is that historical learning can be structured around open, evidence-based questions while preserving explicit teacher guidance. A carefully planned PBL sequence may offer students additional occasions to speak, investigate, and defend their reasoning. Teachers may also need time to develop age-appropriate historical problems and scoring rubrics and to support students less familiar with collaborative inquiry. The source does not provide sufficiently detailed observations to quantify these implementation challenges, so such suggestions are pedagogical implications, not verified descriptions of how often difficulties occurred in the study.

Before publication, the essential revision is an auditable link between the intended critical-thinking construct, its essay items and scoring rubric, the descriptive pretest–posttest table, and each SPSS test. The researchers should recover anonymized student-level results, identify the instrument for each alpha coefficient, and document group assignment, historical problem prompts, instructional fidelity, and any grading procedures. Once the correct outcome is identified, an appropriate baseline-sensitive comparison and uncertainty estimates may be undertaken using real observations. If these materials cannot be recovered, the manuscript should remain a cautiously interpreted descriptive quasi-experimental report rather than claim a definitive impact. Such transparency is more valuable than reporting apparently precise but mismatched statistics.

CONCLUSIONS

The study describes Problem-Based Learning (PBL) in Grade VIII Islamic Cultural History and reports outcome differences between 26 students in the PBL class and 25 in a conventional class. Essay-test means rose from 52.69 to 75.73 in the PBL group and from 46.72 to 54.24 in the comparison group, corresponding to descriptive gains of 23.04 and 7.52 points. The original report also lists normalized gains of .3671 and .2705. However, the independent-samples result, $t(49) = 2.503$, $p = .016$, uses posttest means of 27.54 and 24.56, a scale not reconciled with the essay summaries. It therefore cannot yet confirm the statistical significance of the essay-score difference. Intact-class selection and a baseline gap of 5.97 points further limit causal interpretation. The findings are consistent with promising classroom-level results for PBL, but do not establish the precise effect of the intervention on critical thinking. Verification of primary scores, rubric criteria, instrument identifiers, and actual classroom implementation is necessary before a stronger conclusion is defensible.

REFERENCES

Aba-Oli, Z., Koyas, K., & Husen, A. (2025). Higher-order thinking skills-oriented problem-based learning interventions in mathematics: A systematic literature review. School Science and Mathematics, 125(3), 214–231. https://doi.org/10.1111/ssm.12676

Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. Review of Educational Research, 85(2), 275–314. https://doi.org/10.3102/0034654314551063

Dochy, F., Segers, M., Van den Bossche, P., & Gijbels, D. (2003). Effects of problem-based learning: A meta-analysis. Learning and Instruction, 13(5), 533–568. https://doi.org/10.1016/S0959-4752(02)00025-7

Ennis, R. H. (1991). Critical thinking: A streamlined conception. Teaching Philosophy, 14(1), 5–24.

Ennis, R. H. (2015). Critical thinking: A streamlined conception. In M. Davies & R. Barnett (Eds.), The Palgrave handbook of critical thinking in higher education (pp. 31–47). Palgrave Macmillan. https://doi.org/10.1057/9781137378057_2

Gijbels, D., Dochy, F., Van den Bossche, P., & Segers, M. (2005). Effects of problem-based learning: A meta-analysis from the angle of assessment. Review of Educational Research, 75(1), 27–61. https://doi.org/10.3102/00346543075001027

Harahap, U. K., Selaras, G. H., & Ardi. (2023). Meta-analysis of the impact of problem-based learning on high school students' critical thinking abilities in biology education. Journal of Digital Learning and Education, 3(2), 172–184. https://doi.org/10.52562/jdle.v3i2.751

Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? Educational Psychology Review, 16(3), 235–266. https://doi.org/10.1023/B:EDPR.0000034022.16470.f3

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). Educational Psychologist, 42(2), 99–107. https://doi.org/10.1080/00461520701263368

Kuhn, D. (1999). A developmental model of critical thinking. Educational Researcher, 28(2), 16–25. https://doi.org/10.3102/0013189X028002016

Manuaba, I. B. A. P., No, Y., & Wu, C.-C. (2022). The effectiveness of problem based learning in improving critical thinking, problem-solving and self-directed learning in first-year medical students: A meta-analysis. PLOS ONE, 17(11), e0277339. https://doi.org/10.1371/journal.pone.0277339

Rotgans, J. I., & Schmidt, H. G. (2011). Cognitive engagement in the problem-based learning classroom. Advances in Health Sciences Education, 16, 465–479. https://doi.org/10.1007/s10459-011-9272-9

Rotgans, J. I., O'Grady, G., & Alwis, W. A. M. (2011). Introduction: Studies on the learning process in the one-day, one-problem approach to problem-based learning. Advances in Health Sciences Education, 16, 443–448. https://doi.org/10.1007/s10459-011-9299-y

Rusman. (2018). Model-model pembelajaran: Mengembangkan profesionalisme guru. Rajagrafindo Persada.

Savery, J. R. (2006). Overview of problem-based learning: Definitions and distinctions. Interdisciplinary Journal of Problem-Based Learning, 1(1), 9–20. https://doi.org/10.7771/1541-5015.1002

Schmidt, H. G., Rotgans, J. I., & Yew, E. H. J. (2011). The process of problem-based learning: What works and why. Medical Education, 45(8), 792–806. https://doi.org/10.1111/j.1365-2923.2011.04035.x

Shofa, G. Z., Meiliasari, & Rahayu, W. (2026). The effectiveness of problem-based learning on critical thinking skill: A meta-analysis study reviewed by implementation type, cluster, and geography. Desimal: Jurnal Matematika, 9(2). https://doi.org/10.24042/djm.v9i2.32442

Strobel, J., & van Barneveld, A. (2009). When is PBL more effective? A meta-synthesis of meta-analyses comparing PBL to conventional classrooms. Interdisciplinary Journal of Problem-Based Learning, 3(1). https://doi.org/10.7771/1541-5015.1046

Sugiyono. (2019). Metode penelitian kuantitatif, kualitatif, dan R&D. Alfabeta.

Vecolitha, O., & Yudiarso, A. (2021). Problem based learning (PBL) effectivenessless in improving critical thinking ability: Meta-analysis study. Guidena: Jurnal Ilmu Pendidikan, Psikologi, Bimbingan dan Konseling, 11(2). https://doi.org/10.24127/gdn.v11i2.3716

Walker, A., & Leary, H. (2009). A problem based learning meta analysis: Differences across problem types, implementation types, disciplines, and assessment levels. Interdisciplinary Journal of Problem-Based Learning, 3(1). https://doi.org/10.7771/1541-5015.1061