Jurnal Teknologi Pendidikan, Agustus 2026, 28 (2), 592-605
DOI: http://dx.doi.org/10.21009/JTP2001.6
p-ISSN: 1411-2744 e-ISSN: 2620-3081
Accredited by Directorate General of Strengthening for Research and Development
Available online at http://journal.unj.ac.id/unj/index.php/jtp

AI-Assisted Case Method Learning for Scientific Writing: A Design and Development Study

Nurul Inayah Hutasuhut[1(*)], Reska Mayefis[2], Fitri Yasih[3]
[1,2,3]Universitas Negeri Padang, Padang, Indonesia

Received : June 29, 2026
Revised : July 31, 2026
Accepted : August 15, 2026

Abstract

Undergraduate students are increasingly expected to produce scholarly work, yet writing instruction remains largely lecture-driven and disconnected from students' digital tools. This study designed, validated, and tested instructional products for an undergraduate Scientific Writing course that integrate the case method with Perplexity.ai, an AI answer engine providing source-linked responses. Combining an adapted Recursive Reflective Design and Development (R2D2) model with the Research-Development-Research (RDR) cycle, the study produced lesson plans, thirteen teaching-material units, a four-stage case-method syntax with explicit operational and ethical rules for AI use, and authentic assessment instruments, including a 22-indicator writing rubric (0–100 scale). Three practitioners and three experts judged all five products valid (M = 3.40 on a 4-point scale), and their feedback guided revisions through small-group and large-group trials and a semester-long implementation at Universitas Negeri Padang. Both the model class (n = 21; M = 65.24 to 83.33) and a conventionally taught class (n = 18; M = 62.22 to 72.78) improved significantly (Wilcoxon $p \leq .001$). With classes equivalent at pre-test, analysis of covariance confirmed the model class's advantage, $F(1, 36) = 14.19$, $p = .001$, partial $\eta^2 = .28$, adjusted difference 8.86 points, corroborated non-parametrically ($p = .005$). Students reported high knowledge gains (90.5%) and positive attitudes (81.0%). The findings establish the products' validity and implementability, provide non-randomised evidence that the model outperformed conventional teaching, and show how responsible AI use can become an explicit object of writing instruction.

Keywords: case method; artificial intelligence; Perplexity.ai; scientific writing; design and development

(*) Corresponding Author: inayahytasuhut@fpp.unp.ac.id

How to Cite: Hutasuhut, N. I., Mayefis, R., & Yasih, F. (2026). AI-Assisted Case Method Learning for Scientific Writing: A Design and Development Study. *JTP - Jurnal Teknologi Pendidikan*, *28*(2), 592–605. https://doi.org/10.21009/jtp.v28i2.71033

INTRODUCTION

The capacity to generate original, well-argued scholarly text has become a gatekeeping competence in higher education. Students who cannot organise an argument, marshal evidence, and document sources appropriately struggle not only in writing-intensive courses but also in completing their final projects, which in turn slows time-to-degree and depresses institutional research productivity. Observations in the Indonesian context mirror this pattern: many undergraduates engage only superficially with the composing process, produce papers that restate sources rather than develop an argument, and show little confidence in planning, drafting, and revising scholarly text. These symptoms point less to a deficit in

student ability than to instructional arrangements that fail to demand, model, and support sustained engagement with academic writing.

Two instructional gaps are especially visible in scientific writing courses. First, teaching remains predominantly deductive and lecture-based, casting students as recipients of rules about academic prose rather than as writers working through authentic problems. Second, the learning resources that accompany such courses syllabi, lesson plans, textbooks, and assessment instruments, rarely reflect the technological environment in which students now actually write. Generative and retrieval-based artificial intelligence (AI) tools have moved into students' everyday study practices far faster than curricula have adapted to them (Kasneci et al., 2023; Chiu et al., 2023), even though systematic reviews have long noted how thinly educational perspectives are represented in research on AI in higher education (Zawacki-Richter, Marín, Bond, & Gouverneur, 2019; Chiu, 2024). When course materials ignore these tools, institutions forfeit the opportunity to teach their disciplined, ethical use and instead leave students to improvise, sometimes in ways that threaten academic integrity (Cotton, Cotton, & Shipway, 2024). Recent evidence sharpens the point: heavy, unguided reliance on generative AI among university students has been linked to procrastination and poorer learning outcomes (Abbas, Jam, & Khan, 2024), while students themselves report valuing these tools but wanting clearer institutional guidance on their legitimate use (Chan & Hu, 2023).

The case method offers a pedagogically grounded way to close the first gap. By anchoring instruction in rich, realistic problem situations, it obliges students to analyse, weigh alternatives, and defend decisions, precisely the intellectual moves that scholarly writing demands (Garvin, 2003; Herreid, 2007; Yadav et al., 2007; Krain, 2016). Yet the case method's promise for writing instruction has been only partially realised, because integrating advanced technology into its instructional syntax remains uncommon in everyday teaching practice. Meanwhile, AI answer engines such as Perplexity.ai address the second gap in a distinctive way. The choice of tool was theoretically motivated. Conversational large language models such as ChatGPT generate fluent prose from their internal parameters, so the evidential basis of their answers remains largely opaque to the reader and plausible fabrications can pass unnoticed (Kasneci et al., 2023; Cotton, Cotton, & Shipway, 2024). Perplexity.ai was selected over such chatbots because it is built as a conversational answer engine: each response is composed from retrieved web documents and is inherently linked to explicit, verifiable source citations. This source transparency externalises precisely the behaviours that scientific writing instruction seeks to cultivate locating evidence, judging its credibility, and attributing claims and thereby makes students' information behaviour visible, assessable, and therefore teachable.

This study therefore set out to design, validate, and test a coherent instructional package syllabus, semester lesson plans, teaching materials, an instructional model, and assessment instruments for an undergraduate Scientific Writing course, built around a case-method syntax in which Perplexity.ai functions as a structured inquiry aid. Two questions guided the work: (1) What instructional syntax allows case-method learning of scientific writing to incorporate Perplexity.ai productively and responsibly? (2) Are the resulting products valid according to

practitioners and experts, and implementable in ways that support students' scientific writing development? The study contributes a fully articulated, field-tested design that treats AI not as a shortcut to be policed but as a resource whose boundaries and responsibilities are taught explicitly within the writing curriculum.

METHODS

Research Design

The study followed a design and development research orientation (Richey & Klein, 2007), operationalised by combining two complementary models: an adapted Recursive Reflective Design and Development (R2D2) model (Willis, 1995, 2000) and the Research-Development-Research (RDR) cycle (Gall, Gall, & Borg, 2003). From R2D2 the study took its two working foci definition, and design and development together with the model's recursive, participatory, and non-linear principles; the dissemination focus was excluded because large-scale publication of the products lay beyond the project's scope. From RDR the study took its three activities: a preliminary study, product development, and field implementation of the finished products. The combined procedure therefore comprised (a) a preliminary study, (b) a definition focus, (c) a design and development focus, and (d) a semester-long implementation of the finished products, whose outcomes are reported through descriptive score data, summary-based effect estimates, and qualitative classroom evidence.

Setting and Participants

The research was conducted in the Scientific Writing course of the Department of Family Welfare Science, Faculty of Tourism and Hospitality, Universitas Negeri Padang, Indonesia, during the 2025/2026 academic year. Consistent with R2D2's participatory-team principle, four stakeholder groups took part throughout development. Students, as the products' end users, supplied comments on the materials' appeal, depth, breadth, and developmental appropriateness. Course lecturers collaborated in the design work and reviewed successive drafts. Three practitioners lecturers who regularly teach the Scientific Writing course served as practitioner reviewers. Three experts conducted the expert review: a content expert in scientific writing instruction, an instructional methods expert, and a learning technology expert. Field testing involved a small-group trial with ten students and a subsequent large-group trial with one intact class; each trial ran for approximately three months. In the implementation semester, the intact class taught with the developed products (n = 21) was compared with a parallel class of the same course taught conventionally (n = 18); assignment to classes followed the institution's ordinary enrolment procedures, so the comparison is a non-randomised, intact-group (non-equivalent groups) design.

Development Procedure

The preliminary study gathered information on learning needs, field conditions, and the feasibility of development, and established collaboration with the course lecturers. The definition focus fixed the products to be developed (a)

teaching materials, (b) a course syllabus, (c) semester lesson plans (RPS), (d) an instructional model, and (e) assessment instruments, all oriented to case-method learning supported by Perplexity.ai and constituted the participatory team. In the design and development focus, product drafts were composed collaboratively and then passed through four successive filters: practitioner review, expert review, the small-group trial, and the large-group trial. Reviews used in-depth interviews, discussion, and written appraisal forms covering the five product drafts; revision followed each filter, recursively, until the products were judged sound. The final phase implemented the finished products in the course for one full semester and examined, through observation, document appraisal, and learning journals, how the model functioned in practice and how students' writing engagement and output developed.

Procedures for the Use of Perplexity.ai

Students' use of Perplexity.ai was structured rather than left to improvisation. The tool itself was chosen deliberately: its conversational search-engine architecture inherently links every answer to verifiable web citations, so that information verification-opening, reading, and judging the sources behind an output could be made an explicit, teachable classroom task rather than an act of trust in the machine. Before the case cycle began, students completed a two-session orientation covering how the tool composes answers, how to open and read its source citations, and how to formulate productive queries; the orientation closed with a guided exercise in which each group traced one AI-generated answer back to its listed sources and judged their credibility. During instruction, the tool was used only at the points defined in the syntax concept exploration and concept application and only for three sanctioned functions: locating candidate sources, comparing perspectives on the case problem, and checking disputed claims. A written verification protocol required students to open at least two of the sources cited in any answer they intended to use, confirm that each source actually supported the claim, and record the verification in a worksheet; answers whose citations could not be verified, or that conflicted with the cited sources, were flagged and discussed in class as instances of AI error or hallucination.

The ethical boundaries of AI assistance were made explicit and assessable. Students were prohibited from pasting AI-generated prose into their manuscripts; retrieved material had to be paraphrased, attributed to the original (verified) source, and transformed through the student's own reasoning. Every manuscript was submitted with a short AI-use disclosure statement describing which stages of the work involved Perplexity.ai and for what purpose, in line with recommendations to redefine originality around disclosed and bounded AI assistance (Luo, 2024). Originality was safeguarded through three complementary mechanisms: the portfolio of successive drafts, which made each paper's development visible to the lecturer; a similarity check on the final manuscript; and brief oral questioning during class presentations, in which students explained and defended their sources and arguments. Submitting unedited AI output, citing sources that had not been verified, or omitting the disclosure statement was treated as a violation of the course's academic-integrity norms and handled under the institution's ordinary procedures.

Data Collection and Instruments

Data were of two kinds. Qualitative data comprised the practitioners' and experts' comments, critiques, suggestions, corrections, and appraisals of the products, together with observational records of lecturer and student talk, behaviour, and attitudes during the trials, and the researchers' reflective notes. Quantitative data comprised students' scientific writing scores before and after the implementation semester in both classes, together with descriptive frequencies from students' learning journals. The researchers acted as the principal instrument, supported by (a) a classroom observation guide, (b) an analytic scoring rubric for students' scientific papers, and (c) appraisal guides for the practitioner and expert reviews.

The scoring rubric covered five aspects of scientific writing (1) title and topic selection; (2) development of ideas and content; (3) organisation of ideas and content; (4) presentation technique; and (5) use of scientific Indonesian operationalised in 22 indicators, each rated on a 1–5 scale, yielding a total raw-score range of 22–110; raw totals were converted to a 0–100 scale and rounded to the nearest ten, following the grading convention applied in the course; these tens-based converted scores were the scores archived for each paper and were therefore used both for course grading and for all statistical analyses. No finer-grained (decimal) conversion was retained; the coarseness of the resulting tens-based scale motivated the non-parametric triangulation reported below and is acknowledged among the study's limitations. The rubric was validated in two steps. Its content validity is supported by the practitioner and expert appraisal of the assessment instruments reported in Table 3 ($M = 3.56$, 'very valid'), whose corrections were incorporated before piloting. Item validity was then examined empirically in a tryout with 23 students outside the implementation classes: item–total Pearson correlations exceeded the critical value ($r = .413$; $n = 23$; $p < .05$) for 20 of the 22 indicators, with significant coefficients ranging from .427 to .833, while two indicators (items 19 and 21; $r = .350$ and .266) fell below the criterion and were revised on the basis of the tryout, the final instrument retaining all 22 indicators.

Every pre-test and post-test paper was scored independently by two trained raters the course lecturer and one of the researchers working from the same rubric and anchor examples; discrepancies were resolved through discussion, and the agreed consensus score entered the analysis. A formal inter-rater reliability coefficient could not be computed from the archived consensus scores, which is acknowledged among the study's limitations. The appraisal guides for the practitioner and expert reviews used a four-point rating scale (1 = not appropriate to 4 = very appropriate) covering the content, construct and presentation, and language aspects of each product, with mean ratings of ≥ 3.26 pre-set as the criterion for the 'very valid' category and 2.51–3.25 for 'valid'. Assessment within the model itself was authentic and continuous, using four instruments: scoring rubrics, portfolios, observation sheets, and learning journals.

Data Analysis

Review data were analysed with domain analysis: comments were grouped by content, format, and language domains within each of the five products, reflected upon, and translated into revision decisions; ratings from the appraisal guides were

additionally summarised as means per product and aspect, compared against the validity criteria, and agreement between the practitioner and expert groups was inspected by comparing their mean ratings per product. In addition, the six reviewers' documented verdicts on each of the 20 reviewed components were quantified post hoc as adequate or in need of revision, and content validity was expressed as item-level and scale-level content validity indices (I-CVI and S-CVI/Ave), with I-CVI ≥ .78 as the criterion for six raters (Lynn, 1986). Trial data student and lecturer talk, behaviour, attitudes, and written work were analysed continuously to drive iterative product revision.

Implementation data were analysed on two tracks. Score data were summarised descriptively (means, standard deviations, gains) in SPSS. Given the small class sizes, within-class pre–post change was tested non-parametrically on the converted (0–100) scores with the Wilcoxon signed-ranks test, with the effect size estimated as $r = Z/\sqrt{n}$. The between-class contrast was evaluated with an analysis of covariance (ANCOVA) on post-test scores with pre-test scores as the covariate, so that the comparison is adjusted for initial ability; the homogeneity of regression slopes assumption was checked by testing the pre-test × group interaction, and, given the modest samples and the coarse tens-based score scale, the ANCOVA was triangulated with an independent-samples comparison of gain scores, tested non-parametrically with the Mann–Whitney U test. Unadjusted pooled-variance t-tests with Cohen's d are reported for transparency. Qualitative data observation records, reflective notes, and students' successive drafts were examined for patterns of engagement and composing behaviour, and learning-journal responses were tabulated as frequencies and percentages of the 21 students in the implementation class. Given the small, non-randomised groups, all quantitative results are interpreted with appropriate caution.

RESULTS & DISCUSSION

Findings

The Instructional Model and Product Set

Development yielded an internally consistent product set. The syllabus articulates one competency standard elaborated into thirteen basic competencies for the semester. Semester lesson plans (RPS) operationalise these competencies, each specifying indicators, objectives, materials, methods, learning steps, media, authentic assessment, time allocation, and references. The teaching-material package contains thirteen units organised inductively: each unit opens with worked examples of scholarly text, moves through guided analysis, and closes with production tasks and reflection. The assessment set comprises scoring rubrics, portfolios, observation sheets, and learning journals, with evaluation embedded in the learning tasks rather than appended to them.

The instructional model itself is a four-stage syntax that fuses case-method logic with structured use of Perplexity.ai (Table 1). Across the stages, the lecturer's role shifts from transmitter to facilitator, motivator, and learning partner, while students carry the analytical and compositional work. Perplexity.ai enters at defined points principally concept exploration and concept application where students use

it to locate, compare, and verify sources for the case under analysis and for their own manuscripts. Each use is bounded by explicit rules: retrieved material must be verified against its listed sources, attributed, and transformed by the student's own reasoning; submitting unedited AI output is treated as a violation of the course's academic-integrity norms.

Table 1. Instructional syntax of the case-method model supported by Perplexity.ai

Stage	Core student activity	Role of Perplexity.ai
1. Orientation	Engage with the case; activate prior knowledge of scientific writing concepts; identify key terms and the central problem	Not yet used; the case itself frames the problem
2. Concept exploration	Investigate the concepts the case raises; gather and compare information in small groups	Source-linked searching; students trace and verify the citations attached to each answer
3. Concept interpretation	Negotiate meaning through discussion and debate; present group analyses; test claims against evidence	Consulted to check disputed claims; students must justify acceptance or rejection of retrieved information
4. Concept application	Apply the negotiated concepts to authentic writing tasks (topic selection, problem framing, outlining, drafting, citing, editing)	Idea generation and source discovery for the student's own manuscript, under attribution and originality rules

Practitioner and Expert Validation

The three practitioners and three experts appraised all five products component by component with the four-point appraisal guides, across three aspects: content (accuracy, coverage, and alignment with the competency formulations), construct and presentation (internal consistency, sequencing, layout, and fit with the case-method syntax), and language (clarity, correctness, and appropriateness for undergraduates). Table 3 summarises the ratings: the semester lesson plans (M = 3.53), the teaching materials (M = 3.73), and the assessment instruments (M = 3.56) reached the 'very valid' category (M ≥ 3.26 on the 4-point scale), while the course syllabus (M = 3.04) and the instructional model (M = 3.13) fell in the 'valid' category (2.51–3.25), so that all five products exceeded the minimum feasibility criterion; across all products and both reviewer groups the overall mean was 3.40, and agreement between the two reviewer groups was close, with their mean ratings differing by no more than 0.36 points on any product.

At component level, the first review round covered 20 components across the five products: seventeen were judged adequate by all six reviewers (I-CVI = 1.00); one, the model's comparative advantages, by five of six (I-CVI = .83, above the .78 criterion); and two the implementation guidance, judged insufficiently practical, and the title-selection descriptors of the scoring rubric, two of whose descriptors had been transposed by none (I-CVI = .00). The first-round S-CVI/Ave was therefore .89, with 18 of 20 components meeting the criterion. After the corresponding revisions were executed, the reviewers re-examined the three flagged components and judged all of them adequate, so that every component met the criterion in the second round. Beyond the ratings, most components basic

competencies, materials, success indicators, learning resources, and time allocation were judged appropriate as drafted. Substantive revision requests concentrated on five points; Table 2 pairs each reviewer concern with the specific revision executed in response, so that every change to the products can be traced to the feedback that prompted it. All requested revisions were executed, and the products additionally underwent copy-level correction of typography, terminology consistency, layout, and phrasing after every review and trial round.

Table 2. Principal revision points from practitioner and expert review

Component	Reviewer concern	Revision taken
Syllabus identity	Identity fields left incomplete although the syllabus is intended as a model document	All identity fields completed
Competency standard	Formulation convoluted; core concept insufficiently clear	Statement simplified and conceptually sharpened
Learning process	Constructivist character not yet dominant; lecturer decisions (e.g., material selection) still too prominent	Learning steps re-elaborated to foreground student activity across the four-stage syntax
Evaluation	Mismatch between evaluation procedures and instruments; portfolio scheduled at every meeting judged inappropriate	Instruments aligned with procedures; portfolio repositioned as an end-of-semester holistic assessment
Success indicators / time	Overly detailed indicators risk being ignored in practice; 2 × 50-minute sessions tight for the activity load	Indicators consolidated; activity sequences trimmed to fit the institutional timetable

Table 3. Practitioner and expert appraisal of the five instructional products

Product	Practitioners (M)	Experts (M)	Category
Course syllabus	3.22	2.86	Valid
Semester lesson plans (RPS)	3.42	3.64	Very valid
Teaching materials	3.86	3.60	Very valid
Instructional model	3.26	3.00	Valid
Assessment instruments	3.62	3.50	Very valid

Note. Ratings on a 4-point scale; ‘very valid’ = M ≥ 3.26; ‘valid’ = 2.51–3.25. Cell values are the mean ratings across the content, construct/presentation, and language aspects for each reviewer group.

Field Trials

The two field trials served successive, complementary functions in the development sequence: the small-group trial (ten students, three months) tested the products’ readability, clarity, and workload at close range, while the large-group trial (one intact class, three months) tested whether the full model could be run under ordinary class conditions. Both trials were formative each fed a round of revision and neither involved a comparison class; the comparative element entered only in the subsequent implementation semester. Together, the trials showed that the products could be implemented as designed, aided by the fact that the course lecturers had co-designed them. Trial data nevertheless surfaced four practical lessons. First, residual surface defects typographical errors, unclear sentences, and

occasional mis-sequenced material required correction after each implementation round. Second, students needed explicit early orientation to the model's participation demands, since active knowledge construction was expected from the first meeting. Third, the model is time-hungry: reaching negotiated conclusions through discussion consumed more time than conventional delivery, and students frequently reported that the standard 2 × 50-minute session felt insufficient. Fourth, iterative revision after every meeting, conducted jointly with lecturers and students, was essential to stabilising the products. The trials concluded with a product set judged sound and ready for the implementation semester, which examined the model's feasibility and functioning in regular use and compared its learning gains with those of a conventionally taught parallel class.

Implementation Outcomes

Writing performance. Students' scientific papers were scored before and after the implementation semester in the class taught with the developed products (experimental, n = 21) and in a parallel, conventionally taught class (control, n = 18). Table 4 reports the descriptive statistics. Scores are rubric totals converted to a 0–100 scale. Both classes improved over the semester: the experimental class by 18.10 points (from M = 65.24, SD = 11.23 to M = 83.33, SD = 10.17) and the control class by 10.56 points (from M = 62.22, SD = 11.66 to M = 72.78, SD = 8.95).

Table 4. Descriptive statistics for pre-test and post-test scientific writing scores (0–100 scale)

Group / occasion	n	Min	Max	M	SD	Gain
Experimental, pre-test	21	40	80	65.24	11.23	
Experimental, post-test	21	60	100	83.33	10.17	+18.10
Control, pre-test	18	40	80	62.22	11.66	
Control, post-test	18	60	90	72.78	8.95	+10.56

Wilcoxon signed-ranks tests on the converted scores confirmed that the pre–post improvement was statistically significant in both classes: in the experimental class, $Z = 3.99$, $p < .001$, $r = 0.87$, and in the control class, $Z = 3.42$, $p = .001$, $r = 0.81$, with both effects based on negative ranks (post-test exceeding pre-test). The two classes were statistically equivalent at pre-test, $t(37) = 0.82$, $p = .417$, $d = 0.26$, so the comparison starts from a level baseline. At post-test the experimental class scored 10.56 points higher, $t(37) = 3.41$, $p = .002$, $d = 1.10$. The ANCOVA confirmed this advantage after adjusting for pre-test scores: the effect of group was significant, $F(1, 36) = 14.19$, $p = .001$, partial $\eta^2 = .28$, with baseline-adjusted means of 82.55 (experimental) and 73.69 (control) an adjusted difference of 8.86 points, 95% CI [4.09, 13.63] and the homogeneity of slopes assumption was satisfied (pre-test × group interaction, $F(1, 35) = 0.12$, $p = .733$). The non-parametric triangulation agreed: the experimental class's mean gain of 18.10 points exceeded the control class's 10.56 points, Mann–Whitney $U = 282.5$, $p = .005$, gain-score $d = 0.86$. Taken together, the score data document significant improvement in both conditions and a significantly larger improvement large in standardised terms in

the class taught with the developed model, although the non-randomised assignment and the coarse score scale counsel caution in causal interpretation.

Classroom evidence. The qualitative record differentiates the two conditions more clearly than the scores do. Three patterns were consistent across the semester in the experimental class. First, students' engagement with the full composing cycle broadened markedly: they practised identifying and delimiting topics, formulating titles and problems, constructing thesis statements and outlines, developing paragraphs, handling quotations, drafting, and editing behaviours that had been thin or absent under the previous lecture-based arrangement. Second, the lecturers' appraisal of successive drafts pointed to visible growth in the organisation, argumentation, and source handling of students' papers as the semester progressed. Third, classroom participation rose conspicuously: students were observed to be motivated, enthusiastic, and physically and cognitively active throughout the sequence of case discussions, group presentations, reflections, and independent searches for enrichment material.

Student Engagement and Attitudes

Learning-journal data from the 21 students in the implementation class indicated strong engagement with the model. As shown in Table 5, 13 students (61.9%) reported gaining very substantial knowledge from the course and a further 6 (28.6%) substantial knowledge 90.5% in total while 2 (9.5%) reported limited gains and none reported learning nothing. Attitudinal responses were similarly positive: 5 students (23.8%) reported being very pleased and 12 (57.1%) pleased with the model 81.0% positive in total while 4 (19.0%) were less pleased and none reported displeasure. Observation records corroborated the journals: students participated intensively in discussion, questioning, group presentation, reflection, and independent searching for enrichment material, and took initiative in proposing hypotheses and solutions during case analysis.

Table 5. Students' learning-journal responses during the implementation semester

Dimension	Response category	f (%)
Perceived knowledge gain	Very substantial	13 (61.9%)
	Substantial	6 (28.6%)
	Limited	2 (9.5%)
	None	0 (0.0%)
Attitude toward the model	Very pleased	5 (23.8%)
	Pleased	12 (57.1%)
	Less pleased	4 (19.0%)
	Displeased	0 (0.0%)

Discussion

The findings support the study's central design proposition: that the case method and a source-transparent AI tool can be fused into a single instructional syntax that supports scientific writing development while keeping the student, not the machine, as the author a division of labour consistent with evidence that human–AI collaboration serves learning best when students retain authorial agency (Hwang

& Lee, 2025) and with hybrid human-AI designs in which control is allocated deliberately rather than ceded to the tool (Molenaar, 2022). Four features of the design appear to explain the engagement and writing growth observed during implementation. First, the model repositions students as active constructors of knowledge. Across orientation, exploration, interpretation, and application, studentsnot the lecturer do the analytical work, consistent with constructivist accounts of learning as active meaning-making within a social context (Vygotsky, 1978) and with evidence that participation-centred formats outperform transmission formats for higher-order outcomes (Bonwell & Eison, 1991; Krain, 2016).

Second, the design honours the process nature of writing. The application stage walks students through the full composing cycle planning, translating, and reviewing (Flower & Hayes, 1981) with strategy support of the kind meta-analytic evidence identifies as most effective for adolescent and adult writers (Graham & Perin, 2007). The inductive organisation of the teaching materials, which leads with worked examples before principles, gives students concrete models to reason from rather than abstract rules to memorise.

Third, the integration of Perplexity.ai converts a potential integrity threat into instructional content. Because the tool displays its sources, every retrieval episode doubles as citation practice: students trace claims to references, judge source quality, and decide what to accept habits at the core of scholarly writing. The course's explicit rules on attribution, verification, and originality operationalise the shift that integrity scholarship recommends, from policing AI use to teaching it (Cotton et al., 2024; Kasneci et al., 2023), and give concrete classroom form to recent calls to redefine originality around disclosed and bounded AI assistance (Luo, 2024). In this respect the model functions as a course-level instantiation of the pedagogy dimension of institutional AI-policy frameworks, which assign universities the responsibility of teaching AI literacy rather than merely regulating tool access (Chan, 2023). Students' strongly positive attitudes suggest that they experienced these constraints not as surveillance but as a legitimate part of learning to write.

Fourth, the participatory development process itself contributed to implementability. Because lecturers co-designed the products and reviews were recursive, the versions that reached the effectiveness test had already survived repeated contact with classroom reality an advantage that design and development research explicitly seeks (Richey & Klein, 2007; Willis, 2000). The main friction the trials exposed, time pressure within 2 × 50-minute sessions, is a known cost of discussion-intensive formats and argues for either restructured session blocks or a flipped arrangement in which orientation and initial exploration occur before class.

Three distinct caveats qualify these conclusions. First, although the covariance-adjusted comparison favoured the model significantly and with a large standardised effect, the design was non-randomised: students entered the two intact classes through ordinary enrolment, so unmeasured differences between the classes (lecturer effects, timetabling, group composition) cannot be excluded, and the coarse, tens-based score scale limits the precision of the estimates. The quantitative result is therefore best read as strong preliminary evidence of the model's effectiveness, to be confirmed by randomised or matched replications, rather than as a definitive verdict. Second, an important part of the evidence composing-cycle

engagement, draft quality trajectories, participation is qualitative, and was gathered by researchers who were also the designers. Third, the model's logic presupposes a specific tool affordance source transparency so it transfers to other AI systems only insofar as they make evidence similarly inspectable.

CONCLUSION

This study produced and field-tested a complete instructional package syllabus, lesson plans, thirteen teaching-material units, a four-stage case-method syntax integrating Perplexity.ai under explicit operational and ethical rules, and authentic assessment instruments for undergraduate scientific writing. Practitioner and expert appraisal placed all five products in the 'valid' to 'very valid' categories (overall M = 3.40 on a 4-point scale), and two formative field trials followed by a semester-long implementation indicate that the products are implementable under ordinary course conditions. The implementation yielded statistically significant within-class improvement in writing scores in both the class taught with the products (Wilcoxon $Z = 3.99$, $p < .001$, $r = 0.87$) and the conventionally taught comparison class ($Z = 3.42$, $p = .001$, $r = 0.81$); with the classes equivalent at pre-test, the covariance-adjusted comparison showed a significant advantage for the model, $F(1, 36) = 14.19$, $p = .001$, partial $\eta^2 = .28$, an adjusted difference of 8.86 points on the 0–100 scale. These results establish the validity and implementability of the products and provide non-randomised evidence of the model's effectiveness relative to conventional teaching, alongside consistent qualitative evidence of student growth: broad engagement with the full composing cycle, visible development in the organisation and source handling of students' papers, high self-reported knowledge gain (90.5% substantial or very substantial), and predominantly positive attitudes (81.0%). Because assignment to classes was not randomised, this advantage is best confirmed through randomised or matched replications. The study's broader contribution is a worked demonstration that responsible AI use can be designed into a writing course's instructional model rather than bolted on as policy.

The findings are bounded by a single course, a single institution, small non-randomised intact classes ($n = 21$ and $n = 18$), and a coarse 0–100 score scale finalised by rater consensus without formal reliability coefficients; the dissemination phase of development was not undertaken. Future research should replicate the model in other courses and institutions with samples sized to detect a moderate effect, use randomised or matched designs that secure baseline equivalence, report formal inter-rater and internal-consistency reliability, examine longer-term retention of writing gains, and test whether the syntax generalises to AI tools with different transparency affordances.

ACKNOWLEDGEMENT

This research was funded by the internal grant of the Institute for Research and Community Service, Universitas Negeri Padang, fiscal year 2025 [Contract no:

2024/UN35.15/LT/2025]. The authors thank the practitioner and expert reviewers, the course lecturers, and the students of the Scientific Writing course for their participation and feedback throughout the development and implementation of the products.

REFERENCES

Abbas, M., Jam, F. A., & Khan, T. I. (2024). Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. *International Journal of Educational Technology in Higher Education, 21,* Article 10.

Bonwell, C. C., & Eison, J. A. (1991). *Active learning: Creating excitement in the classroom* (ASHE-ERIC Higher Education Report No. 1). The George Washington University.

Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education, 20,* Article 38.

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education, 20,* Article 43.

Chiu, T. K. F. (2024). Future research recommendations for transforming higher education with generative AI. *Computers and Education: Artificial Intelligence, 6,* 100197.

Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence, 4,* 100118.

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International, 61*(2), 228–239.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365–387.

Gall, M. D., Gall, J. P., & Borg, W. R. (2003). *Educational research: An introduction* (7th ed.). Allyn & Bacon.

Garvin, D. A. (2003). Making the case: Professional education for the world of practice. *Harvard Magazine, 106*(1), 56–65.

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99*(3), 445–476.

Herreid, C. F. (2007). *Start with a story: The case study method of teaching college science.* NSTA Press.

Hwang, Y., & Lee, J. H. (2025). Exploring students' experiences and perceptions of human–AI collaboration in digital content making. *International Journal of Educational Technology in Higher Education, 22,* Article 44.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103,* 102274.

Krain, M. (2016). Putting the learning in case learning? The effects of case-based approaches on student knowledge, attitudes, and engagement. *Journal on Excellence in College Teaching, 27*(2), 131–153.

Luo, J. (2024). A critical review of GenAI policies in higher education assessment: A call to reconsider the 'originality' of students' work. *Assessment & Evaluation in Higher Education, 49*(5), 651–664.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*(6), 382–385.

Molenaar, I. (2022). Towards hybrid human–AI learning technologies. *European Journal of Education, 57*(4), 632–645.

Richey, R. C., & Klein, J. D. (2007). *Design and development research: Methods, strategies, and issues.* Lawrence Erlbaum Associates.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Harvard University Press.

Willis, J. (1995). A recursive, reflective instructional design model based on constructivist-interpretivist theory. *Educational Technology, 35*(6), 5–23.

Willis, J. (2000). The maturing of constructivist instructional design: Some basic principles that can guide practice. *Educational Technology, 40*(1), 5–16.

Yadav, A., Lundeberg, M., DeSchryver, M., Dirkin, K., Schiller, N. A., Maier, K., & Herreid, C. F. (2007). Teaching science with case studies: A national survey of faculty perceptions of the benefits and challenges of using cases. *Journal of College Science Teaching, 37*(1), 34–38.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: Where are the educators? *International Journal of Educational Technology in Higher Education, 16,* Article 39.