**Jupiter**: Jurnal Pendidikan Terapan
Homepage : https://journal.diginus.id/jupiter/index
Email: jurnalpendidikanterapan@gmail.com

p-ISSN : 2985-3214; e-ISSN : 2964-3171
**Jupiter**, Vol. 4, No. 5, November 2026
Page 161-174

# From Task Completion to Claim Calibration: Cycle 1 Design-Based Research on Scientific Writing Scaffolding

[1*]**Muhammad Syaiful Nurasman, [2]Sri Nur Yuliyawati, [3]Hazma, [4]Tedy Heriyadi**
[1234]Politeknik Negeri Bandung, Bandung, Indonesia

**ARTICLE INFO**

**Article history:**
Submitted: July 10, 2026
Final Revised: Sept. 18, 2016
Accepted: October 06, 2026
Published: November 06, 2026

**Keywords:**
Design-Based Research; Scientific Writing; Vocational Higher Education; Instructional Scaffolding; Claim Calibration

**ABSTRACT**

Scientific writing is a core academic competency in vocational higher education, yet instructional support remains inconsistent in Indonesia's Diploma-4 (D4) programmes, which favour applied and technical competencies. This paper reports Cycle 1 of a Design-Based Research project that aimed to design, implement, and evaluate a six-stage scaffolding workflow (P1–P6) embedded in the Indonesian Language course across three D4 programmes (accounting, engineering, and tourism) at a single vocational polytechnic in Indonesia. Data comprised 26 fully coded artefacts from six purposively selected focal groups and one additional stage-specific comparative artefact, field notes, and a reflection questionnaire completed by 72 of 180 enrolled students (40% response rate), analysed through qualitative content analysis supported by peer debriefing with one independent reviewer. The scaffold supported students through the structural sequence of mini-research writing but did not yet adequately scaffold claim calibration, variable operationalisation, and peer evaluation. Premature causal, correlational, and significance claims appeared in 12 of 26 coded artefacts (46%), concentrated at the interpretation stage, while indicator-item misalignment emerged as a persistent operationalisation bottleneck. Students rated the initial framing stage as least demanding, yet artefact analysis suggested it may be analytically the most consequential for later claim-calibration difficulties, tentatively termed the P1 paradox. Perceptions data suggest peer-review quality may have been constrained by inadequate feedback architecture, though the data cannot rule out a competence deficit as a contributing explanation. Five tentative design principles for Cycle 2 are proposed, targeting claim calibration, phenomenon framing, indicator traceability, peer-review criteria, and worked-example fading.

**Doi:** https://doi.org/10.61255/jupiter.v4i5.1592

## INTRODUCTION

Scientific writing ("*karya tulis ilmiah*") is a core academic competency in Indonesian higher education, involving problem formulation, methodological design, data collection, analysis, and evidence-based discussion. Its integration into vocational programmes remains inconsistent and under-theorised internationally, not only in Indonesia, as writing pedagogy in vocational and post-compulsory disciplines remains comparatively under-examined relative to academic-track instruction (Stubley & Papen, 2026). Studies of Indonesian D4 vocational writing instruction report a persistent misalignment between course design, outcome-based education orientation, and explicit support for empirical writing (Hidayat et al., 2023; Putera & Shofiah, 2021; Wienanda et al., 2021), while a recent multi-programme study found that higher-order writing competencies synthesis, critical evaluation, and evidence-based argumentation remain underdeveloped, with students treating scientific writing chiefly as structural compliance rather than disciplined inquiry (Putera et al., 2025).

This reflects a deeper gap: within the Indonesian D4 vocational writing literature reviewed above, we did not identify a structured, sequential instructional design guiding students through the full epistemic arc of inquiry-based writing, from identifying a researchable phenomenon to the reasoned communication of findings. Genre-based pedagogy holds that writing competence develops through guided participation in discipline-relevant tasks rather than exposure to abstract rules (Hyland, 2007). The compulsory, cross-disciplinary Indonesian Language course (Republik Indonesia, 2012) is well positioned to serve this role, yet its potential to develop disciplinary reasoning and evidence-

based communication (Bassett & Macnaught, 2025; Putera et al., 2025) remains largely untapped. Macro-level task-sequence support is known to improve metacognitive outcomes in language and literacy instruction (Hammond & Gibbons, 2005), and a similar pattern has been observed more broadly in online higher-education scaffolding (Doo et al., 2020); however, neither line of evidence speaks specifically to writing pedagogy, and task-sequencing support alone appears insufficient for scientific writing, a gap also noted in project-based academic writing instruction research (Nurasman et al., 2026): students also need support to monitor whether claims are warranted by evidence (Anggraeni et al., 2025; Taub et al., 2023) and to develop the feedback literacy peer review requires (Carless & Boud, 2018; Malecka et al., 2022). Indonesian studies to date have concentrated on curriculum orientation and needs analysis (Hidayat et al., 2023; Wienanda et al., 2021), with little examination of a systematically designed, sequential scaffold targeting these competencies. Where project-based interventions have been tested in Indonesian vocational settings, they have targeted generic 4Cs competencies: critical thinking, communication, collaboration, and creativity rather than the specific epistemic-writing competencies this study targets (Syahril et al., 2022). What distinguishes the workflow reported here is not any single scaffolding mechanism in isolation, since task-sequencing, evidence-monitoring, and feedback-literacy supports have each been studied separately, but their integration into one sequential design spanning phenomenon framing, operationalisation, data collection, and peer feedback within a single compulsory course, implemented and evaluated as a coherent system rather than as isolated interventions.

Design-Based Research (DBR) is well suited to this gap: it develops functional instructional solutions and generalisable design principles through iterative cycles of design, implementation, and evaluation in authentic settings (Anderson & Shattuck, 2012; Hoadley & Campos, 2022; McKenney & Reeves, 2018), treating the practitioner-researcher as an active design agent. This paper reports Cycle 1 of a DBR project that designed, implemented, and analysed a six-stage sequential scaffolding workflow (P1–P6) embedded in the Indonesian Language course across three D4 vocational study programmes (Accounting, Energy Conservation Engineering, and Tourism Destination Management) at an Indonesian polytechnic.

Three research questions guide the study: (1) how was the P1–P6 scaffolding workflow designed and implemented across the three study programmes; (2) what conceptual writing patterns emerged across the workflow's artefact-producing stages; and (3) what design principles for Cycle 2 can be derived from these Cycle 1 findings. The study aims to contribute both a practical instructional design and theoretically grounded design principles to the literature on scientific writing pedagogy in Indonesian vocational higher education.

## METHOD

### Research Design

This study employs Design-Based Research (DBR) (Hoadley & Campos, 2022; McKenney & Reeves, 2018), reporting Cycle 1 of a planned two-cycle project comprising design and construction of the P1–P6 framework and its worksheets, implementation across three study programmes in one semester, and evaluation through artefact-based analysis, intended to directly inform the Cycle 2 redesign.

### Context and Participants

The study was conducted in the compulsory, cross-disciplinary Indonesian Language course (Republik Indonesia, 2012) at a D4 vocational polytechnic in Bandung, Indonesia, across three study programmes: Accounting (AC), Energy Conservation Engineering (TKE), and Tourism Destination Management (DP). The researcher served as lecturer for all three programmes, fulfilling the dual practitioner-researcher role characteristic of DBR; this positionality was addressed through an independent peer reviewer in the artefact-analysis phase (see Data Analysis below).

The P1–P6 Scaffolding Framework. The framework (Figure 1) scaffolds a mini-research project through six sequential stages, each supported by a structured worksheet (Lembar Kerja, LK) that doubled as a scaffolding tool and data artefact: P1, phenomenon identification and research-question formulation; P2, theoretical review and variable operationalisation into indicators and instrument items; P3, methodology and data-tabulation design; P4, data collection, tabulation, and preliminary interpretation; P5, structured peer review and revision (documented via field notes); and P6, final report submission. Stages ran from Week 9 across six class meetings, supported by complementary activities: Socratic probing at P1, a gap-evidence checklist at P2, a data-to-interpretation traceability table at P4, and comparative benchmarking at P5. One in-cycle adjustment, relocating preliminary analysis from P3 to P4 after observed cognitive overload, was recorded as implementation evidence.

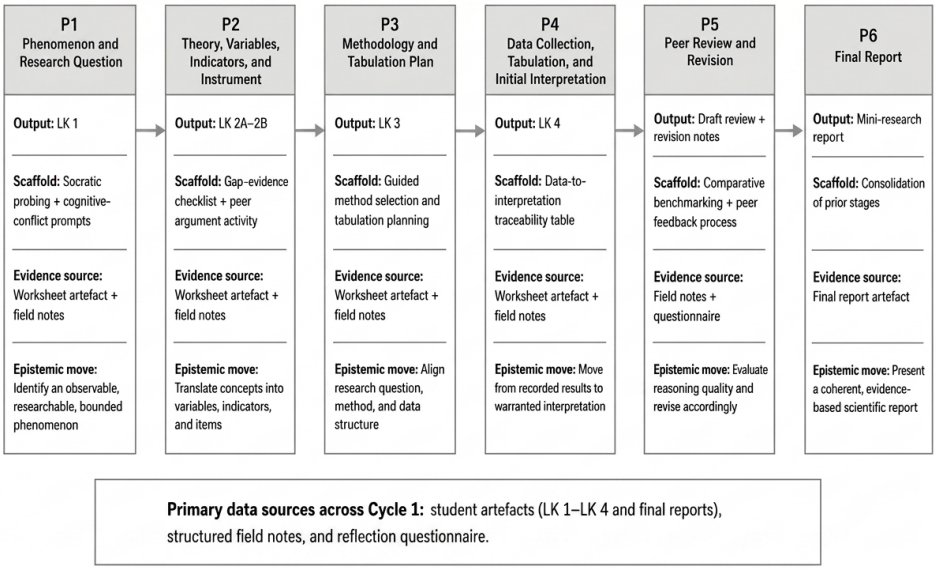



**Figure 1.** P1–P6 sequential scaffolding workflow for mini-research scientific writing in vocational higher education.

**Data Collection**

Three complementary sources were used: student learning artefacts (LK 1–4 and final reports) as the main source; structured observational field notes completed at each session; and a reflection questionnaire administered after P1–P5 completion as a triangulation source for perceived difficulty and usefulness rather than a psychometric measure. From each of the three participating study programmes, two groups were selected through criterion-based purposive sampling, yielding six analysed groups. Within each programme, the anchor case was selected against two pre-analysis criteria: a complete longitudinal artefact trail (LK1–LK4 available and legible, plus a submitted final report, enabling the trajectory from initial framing to final product to be traced) and continuity of inquiry (the phenomenon and research question formulated at P1 remained traceable through to the final report, with refinement permitted but not substitution of the underlying project). A second, additional case was included in each programme to broaden artefact variation beyond the longitudinal anchor case, so that the coded corpus was not restricted to a single trajectory per programme. Selection was made before formal coding and was not based on rubric standing or on the presence or absence of any analytical code; across the six selected groups, final-report rubric scores spanned 16–22/24. Variation in reasoning quality, implementation constraints, and artefact form observed within the selected cases is reported as an analytical finding below, not as a basis for their selection. The resulting documentation corpus comprised 30 learning artefact sets: 24 staged worksheet submissions and six final research reports (P6). Twenty-three of the staged submissions came from the six focal groups, one of which did not produce an LK4; one further LK4, from a DP group outside the focal set, was retained as stage-specific comparative evidence at P4. Of these, 26 artefacts met the inclusion criteria of sufficient completeness, readability, and relevance to the analytical focus and were fully coded. Four artefacts one each at P2 and P4, and two at the final-report stage, did not meet these structural criteria and were retained as contextual documentation rather than formally coded. This screening was conducted prior to coding and was based on completeness and legibility rather than content, and no artefact was excluded on the basis of its substantive findings or coding outcome (the resulting per-stage denominators are reported in Table 2). The study involved 180 enrolled students across the three participating study programmes, organised into 40 groups. Six groups were purposively selected for the artefact-based analysis, while the reflection questionnaire was administered to all enrolled students; 72 students provided valid responses (40.0% response rate; Table 1).

**Table 1.** Participant and Group Summary by Study Programme

| Study Programme | Students Enrolled | Groups Formed | Groups Analysed |
|---|---|---|---|
| Accounting (AC) | 59 (2 classes) | 12 | 2 |
| Energy Conservation Engineering (TKE) | 60 (2 classes) | 14 | 2 |
| Tourism Destination Management (DP) | 61 (2 classes) | 14 | 2 |
| **Total** | **180** | **40** | **6** |

*Note. Group size ranged from three to five students, depending on programme-level enrolment and class organisation. The questionnaire was administered to all 180 enrolled students, yielding 72 valid responses overall (40.0% response rate).*

**Questionnaire Development**

Before administration, the items were reviewed internally by two co-investigators with relevant expertise in Indonesian language pedagogy and academic writing. The review considered item relevance to the intended indicators, clarity of wording, and alignment with the instructional sequence. No pilot test was conducted; this is acknowledged as a limitation, and questionnaire findings are accordingly interpreted as descriptive indicators of student perception rather than as evidence from a psychometrically validated instrument. The 40% questionnaire response rate raises the possibility of nonresponse bias; however, because the questionnaire was administered anonymously and no information was collected from nonrespondents, differences between respondents and nonrespondents could not be assessed.

**Data Analysis**

Artefact analysis followed qualitative content analysis adapted for DBR evaluation (McKenney & Reeves, 2018), using an eight-aspect analytic reading framework (Appendix A) as an audit-trail instrument. Each aspect was scored on a 0–3 scale (3 = strong, 2 = adequate, 1 = needs strengthening, 0 = not available), yielding a maximum total of 24; this is the same instrument underlying the final-report scores reported below, and constitutes an analytic research score rather than students' official course grade. The codebook comprised four non-mutually-exclusive categories: M (misconception in reasoning or claim, with subcodes M1a/M1b/M3a–d/M4), K (conceptual gap), R (register deviation), and L (positive writing behaviour); full definitions appear in Appendix A. The unit of coding was the individual claim, phrase, or sentence exhibiting a codable pattern; the unit of analysis reported in Table 2 is the artefact as a whole, coded as present or absent for each category based on whether at least one qualifying instance was identified within it. Code assignment followed a primarily inductive logic: candidate patterns were first identified in the artefact text and then matched against the codebook's boundary criteria (Appendix A), with the codebook itself refined iteratively as new instances were encountered. Where an artefact exhibited more than one applicable code, all were retained rather than resolved to a single dominant label, consistent with the non-mutually-exclusive scheme reported in Table 2. Because the codebook was developed iteratively during coding rather than fixed in advance, it carries an inherent risk of confirmation bias, where emerging codes are shaped by the patterns the researcher is already documenting; this risk was addressed, though not eliminated, through the eight-aspect audit-trail instrument above, which records the interpretive basis for each code assignment, and through the peer debriefing described next. Credibility was strengthened through peer debriefing with an independent Indonesian Language lecturer, who reviewed eight purposively selected artefacts spanning the full range of coding patterns; seven were judged defensible as coded, and one (DP Group 4 final report) generated a documented consensus decision to add an R code alongside the existing M3c code, reflecting a broader mechanism of stylistic mimicry incorporated into the Discussion below. Field notes were analysed thematically; questionnaire responses were analysed descriptively (item means) and open-ended responses thematically coded by stage.

**Trustworthiness and Ethics**

Trustworthiness was supported through prolonged engagement, peer debriefing, methodological triangulation, negative-case analysis, and a detailed methodological log. The instructional activities examined in this study were conducted as part of regular coursework and were not introduced solely for research purposes. The learning artefacts analysed were produced through students' ordinary coursework rather than through instruments designed exclusively for data collection. Before the P1–P6 sequence, students were informed in class that their learning artefacts might be used for research purposes and that their decision to participate in the research would not affect their course grades. The reflection questionnaire was administered anonymously through an online form (Google Forms) on a voluntary basis, with no identifying information collected and no obligation to respond. All groups and quotations are reported using anonymised identifiers (e.g., AC Group 1) to protect participant confidentiality.

Because Cycle 1 was conducted within regular, non-experimental coursework, formal institutional ethics review was not obtained at the time of implementation. This constitutes a limitation of the study, particularly given the dual role of the lecturer as instructor and researcher. The study therefore reports the procedures used to protect voluntariness and confidentiality, while formal institutional ethics clearance will be sought for Cycle 2 before the collection and analysis of research data.

## RESULTS

This section reports the implementation of the P1–P6 workflow, the cross-programme artefact-coding patterns from 26 fully coded artefacts, and the questionnaire findings from 72 students.

**Implementation Overview**

The P1–P6 workflow ran across six class sessions between Weeks 9 and 15 in all three programmes; all stages were completed, though implementation varied in analytically relevant ways. P1 and P2 proceeded smoothly, though groups often reached topic consensus quickly without weighing whether a phenomenon was observable and answerable, and bridging the gap-evidence checklist (LK 2A) to instrument design (LK 2B) required explicit instructor support. P3 was the most cognitively demanding stage; preliminary data analysis, originally planned for P3, was relocated to P4 after observed cognitive overload, an in-cycle adjustment recorded as a Cycle 1 finding. The P4 data-to-interpretation traceability table served as an effective anchor, though claim calibration remained problematic. P5 peer review was completed by all groups, but field notes recorded predominantly surface-level comments absent a structured instrument; the comparative-benchmark activity was its most effective element. All six groups submitted final reports (rubric scores 16–22/24); one DP session, affected by a scheduling conflict, was resolved via take-home submission without apparent quality loss.

**Artefact Coding Patterns**

Of the 30 assembled artefact sets, 26 met inclusion criteria and were fully coded across the P1–P4 and P6 stages (P5 is documented through field notes). Table 2 presents the frequency and percentage of each code by stage.

**Table 2.** Frequency of Coded Patterns by Stage (N = 26 Artefacts)

| Stage | n | M1b | M3a | M3b | M3c | M4 | M3 (any subtype) |
|---|---|---|---|---|---|---|---|
| P1 | 6 | 3 (50%) | 2 (33%) | 0 (0%) | 0 (0%) | 0 (0%) | 2 (33%) |
| P2 | 5 | 0 (0%) | 0 (0%) | 3 (60%) | 0 (0%) | 0 (0%) | 3 (60%) |
| P3 | 6 | 1 (17%) | 0 (0%) | 0 (0%) | 0 (0%) | 3 (50%) | 0 (0%) |
| P4 | 5 | 0 (0%) | 3 (60%) | 0 (0%) | 2 (40%) | 0 (0%) | 4 (80%) |
| P6 (Final Report) | 4 | 0 (0%) | 2 (50%) | 0 (0%) | 2 (50%) | 0 (0%) | 3 (75%) |
| **Total** | **26** | **4 (15%)** | **7 (27%)** | **3 (12%)** | **4 (15%)** | **3 (12%)** | **12 (46%)** |

*Note. M = misconception in reasoning or claim; M1b = overscoped phenomenon or research question; M3a = premature causal claim; M3b = claim implied beyond instrument capacity; M3c = significance asserted without an inferential basis; M4 = indicator–item misalignment; R = register deviation. Codes are non-mutually exclusive; an artefact may receive more than one code, so row totals may exceed n. Denominators vary by stage (n = 4–6) because four artefacts were excluded from the coded corpus (see Data Collection); percentages are calculated against the coded artefacts at each stage, not the full six-group sample. As artefacts are nested within a small number of purposively selected groups rather than representing independent individual students, these figures describe the distribution of coded artefacts, not the prevalence of the underlying pattern among individual students. P5 is discussed narratively (see Implementation Overview and Peer Review sections) rather than included here. Positive contextual variations are discussed narratively as contrasting cases rather than reported as formal coded patterns.*

M3 (premature causal, correlational, and significance claims) appeared from P1 onward, consolidating at P4 (80% of coded artefacts at that stage) and persisting into selected final reports (75%), while M4 (indicator–item misalignment) emerged as the dominant operationalisation pattern at P3, where M3 was entirely absent. At P1, M3 was confined to AC: two groups framed their research questions using causal diction ("pengaruh", "meningkatkan") despite a descriptive design, whereas TKE and DP artefacts at the same stage showed M1b (overscoped phenomena) instead. M3 was present in 12 of the 26 coded artefacts (46%), a figure that describes the distribution of artefacts across six nested groups rather than an estimate of individual-student prevalence; the remaining 14 (54%) were M3-free. M3 did not manifest identically across programmes: it appeared as causal verbs applied to descriptive survey data in AC, as a combined M3a+M3c pattern across stages in TKE's anchor group, and as a progression from M3b to M3c (and M3c+R) in DP. Nor was it fixed to individual groups: the AC Group 2 anchor case shows M3a appearing at P1, receding at P3 where structural concerns dominated, re-emerging at P4, and diminishing again in the final report.

The negative cases are equally instructive. Artefacts free of M3 shared three features: bounded research questions, descriptive rather than inferential findings language (e.g., *sebagian besar responden menyatakan* rather than *terbukti bahwa*), and discussion sections that treated theory as an interpretive lens rather than a causal mechanism. The clearest case, DP Group 1, was M3-free from P1 through P4; its topic (an observable field phenomenon at a tourism destination) structurally limited opportunities for inferential overreach. This group's peer-review session was recorded in field notes as the least substantively productive of all observed sessions.

What makes M3 the central design problem of Cycle 1 is not that students erred, but that this specific error persisted even in structurally strong work. TKE Group 7's final report received a rubric score of 22/24 and one of the most systematic data-documentation sets in the corpus, yet its abstract concluded that instrument and laboratory facility inadequacy was proven to correlate negatively and significantly with learning time efficiency, a claim its own methods section had already limited to descriptive statistics (minimum, maximum, mean).

**Student Perceptions: Workload, Scaffolding, and Stage Difficulty**

The reflection questionnaire, completed by 72 students (AC n = 33; TKE n = 22; DP n = 17), assessed workload, learning value, workflow coherence, and scaffolding effectiveness on 1–5 Likert items (Table 3).

**Table 3.** Student Questionnaire Mean Scores by Study Programme (n = 72)

| Questionnaire item | AC (n=33) | TKE (n=22) | DP (n=17) | Overall (n=72) |
|---|---|---|---|---|
| ***Panel A. Perceived workload by stage (1 = very light; 5 = very heavy)*** | | | | |
| P1: Scientific claim and introduction | 2.85 (0.97) | 2.86 (0.71) | 3.12 (0.78) | 2.92 (0.85) |
| P2: Instrument design | 3.45 (0.94) | 3.23 (1.07) | 3.53 (0.80) | 3.40 (0.94) |
| P3: Methodology, theoretical review, and tabulation | 3.39 (1.03) | 3.18 (1.01) | 3.71 (0.69) | 3.40 (0.96) |
| P4: Findings and discussion | 3.79 (0.99) | 3.59 (0.80) | 3.65 (0.79) | 3.69 (0.88) |
| P5: Peer review and revision | 3.33 (0.92) | 3.36 (0.85) | 3.35 (0.93) | 3.35 (0.89) |
| ***Panel B. Perceived learning value by stage (1 = not beneficial; 5 = very beneficial)*** | | | | |
| P1: Scientific claim and introduction | 4.52 (0.76) | 4.36 (0.79) | 4.41 (0.51) | 4.44 (0.71) |
| P2: Instrument design | 4.58 (0.66) | 4.41 (0.73) | 4.18 (0.73) | 4.43 (0.71) |
| P3: Methodology, theoretical review, and tabulation | 4.55 (0.67) | 4.36 (0.90) | 4.47 (0.62) | 4.47 (0.73) |
| P4: Findings and discussion | 4.64 (0.65) | 4.36 (0.85) | 4.35 (0.70) | 4.49 (0.73) |
| P5: Peer review and revision | 4.52 (0.76) | 4.45 (0.80) | 4.29 (0.85) | 4.44 (0.79) |
| ***Panel C. Workflow sequence and LK scaffolding (1–5 Likert scale)*** | | | | |
| Sequence P1–P5 felt logical and coherent | 4.58 (0.56) | 4.05 (0.90) | 3.94 (0.75) | 4.26 (0.77) |
| Each stage prepared me for the next | 4.52 (0.57) | 4.14 (0.89) | 3.94 (0.75) | 4.26 (0.75) |
| I knew the purpose of each stage while working | 4.33 (0.69) | 4.05 (0.72) | 3.94 (0.66) | 4.15 (0.71) |
| Examples in the LK helped my understanding | 4.64 (0.55) | 4.23 (0.87) | 4.29 (0.69) | 4.43 (0.71) |
| The LK column structure aided my writing | 4.42 (0.75) | 4.14 (0.89) | 4.18 (0.73) | 4.28 (0.79) |
| LK 3 was manageable in one session | 4.06 (0.83) | 3.86 (0.99) | 3.76 (0.66) | 3.93 (0.84) |
| LK 4 helped distinguish recording from interpreting data | 4.21 (0.74) | 3.91 (0.87) | 3.82 (0.73) | 4.03 (0.79) |
| ***Panel D. Peer review (1–5 Likert scale)*** | | | | |
| Peer review helped me see weaknesses in my writing | 4.33 (0.65) | 3.95 (0.65) | 3.94 (0.83) | 4.12 (0.71) |
| Comments from other groups were specific enough to act on | 3.39 (0.90) | 3.77 (0.87) | 3.53 (0.80) | 3.54 (0.87) |
| I made meaningful revisions after peer review | 4.18 (0.85) | 4.05 (0.72) | 4.06 (0.83) | 4.11 (0.80) |
| ***Panel E. Support and group collaboration (1–5 Likert scale)*** | | | | |
| I felt adequately supported (by LK, examples, and instructor guidance) | 4.45 (0.67) | 4.23 (0.81) | 4.18 (0.81) | 4.32 (0.75) |
| Group work in this project was collaborative and fair — workload was evenly shared | 4.06 (1.00) | 3.86 (0.89) | 3.47 (1.18) | 3.86 (1.03) |

*Note. All items were rated on a 1–5 scale. Overall means were calculated from all valid respondent-level responses (n = 72), rather than by averaging programme-level means. P1–P5 refer to the student-facing task descriptions used in the questionnaire instrument itself, which differ in wording from the analytical process labels used elsewhere in this article (see Figure 1); in particular, the questionnaire's P3 item bundles theoretical review with methodology, whereas the analytical framework assigns theoretical review to P2, and the questionnaire's P4 item asks about findings and discussion broadly, corresponding to data collection through preliminary interpretation in the analytical framework. The questionnaire's P1 item likewise uses student-facing wording ('Scientific claim and introduction') for the stage the analytical framework defines as phenomenon identification and research-question formulation; comparisons between P1 perceptions and P1 artefact patterns should be read with this difference in wording in mind.*

P4 (findings and discussion) was rated the most demanding stage overall (3.69/5; AC highest at 3.79), directly aligning with the artefact-level concentration of M3 at this stage; P4 also received the highest learning-value score (4.49/5).

Perceptions of scaffolding quality were generally positive: 'Examples in LK helped my understanding' received the highest mean among the scaffolding items (4.43/5), and workflow coherence items ('Sequence P1–P5 felt logical', 'Each stage prepared me for the next') both scored 4.26/5. Two items scored lower: 'LK 3 was manageable in one session' received the lowest LK-related mean (3.93/5; DP highest burden at 3.76), attributed to difficulty locating and

synthesising theoretical sources within a single session; and 'LK 4 helped distinguish recording from interpreting data' scored 4.03/5, with TKE (3.91) and DP (3.82) lower.

Open-ended responses corroborate the artefact patterns: writing the research gap was the most frequently cited P1 difficulty (37 of 72 respondents), followed by identifying a suitable phenomenon (34) and formulating the research question (23). At P2–P3, 35 respondents reported difficulty writing questionnaire items and 32 reported difficulty determining indicators. At P4, 74% of respondents (53 of 72) cited difficulty connecting findings to theory, the single highest-frequency theme in the questionnaire.

Peer review (P5) produced the sharpest internal divergence in the questionnaire data: while 'Peer review helped me see weaknesses in my writing' (4.12/5) and 'I made meaningful revisions after peer review' (4.11/5) were both rated positively, 'Comments from other groups were specific enough to act on' received the lowest mean of any evaluative item in the questionnaire (3.54/5; AC lowest at 3.39). Field observation recorded that, without a structured review instrument, reviewers defaulted to surface-level commentary, and in several groups students were still completing their own drafts during the peer-review session.

## DISCUSSION

Although participants were drawn from three study programmes, the purpose was not to compare disciplinary performance; the programmes functioned as contrasting implementation contexts for examining the robustness of the P1–P6 design, so the discussion foregrounds cross-programme patterns rather than between-programme comparisons.

### From Task Completion to Claim Calibration: M3 as the Central Design Problem

The persistence of M3 even in structurally strong work, illustrated by TKE Group 7's report, indicates that the gap between analytical method and conclusion strength was not a product of incomplete work but of a scaffold that had not made that distance visible: the P1–P6 design sequenced tasks and supplied templates but never posed the question of what kind of claim this kind of data licenses. This distinction between task completion and claim calibration is the central design problem this study identifies, consistent with research on self-regulated writing that stresses monitoring whether emerging claims are evidentially warranted, not only completing the task (Taub et al., 2023), and with claim calibration as itself a linguistic competency requiring writers to tune certainty through modal verbs, reporting verbs, and stance adverbials proportionate to their evidence (Lingard, 2020). For Cycle 2, this points toward an explicit claim–analysis alignment checkpoint at P4, mapping analytical procedure to the verbs it licenses (descriptive evidence to *menunjukkan/mengindikasikan*; associative evidence to *berhubungan/berkaitan;* causal evidence only to *memengaruhi/menyebabkan*), formalised as Design Principle 1 in Table 4.

A second, more tentative interpretive layer emerged from peer debriefing of the DP Group 4 final report (coded M3c+R): some M3 instances may reflect stylistic mimicry, the adoption of inferential academic register *('signifikan', 'berpengaruh', 'korelasi', 'terbukti'*) without the methodological apparatus that gives these terms meaning in their source texts, rather than a reasoning error as such. This would suggest a language-socialisation process, students acquiring the surface features of scientific register from exposure to journal articles and theses, consistent with evidence that novice L2 writers can struggle to deploy evidential hedges proportionate to their evidence (Barbara et al., 2024). The interpretation remains tentative, grounded in a small number of cases, but it carries a distinct design implication: explicit instruction in epistemic hedging language, not only procedural scaffolding, an approach well suited to embedded academic-literacy instruction (Bassett & Macnaught, 2025). The negative cases indicate that M3 is partly associated with the quality of early framing, and that the most efficient point of Cycle 2 intervention may lie upstream of P4, at P1 itself, a divergence taken up further below.

### The P1 Paradox: Early Framing and Later Claims

Students rated P1 as the least demanding stage of the entire workflow (mean 2.92/5), while artefact analysis suggests it may be analytically among the most consequential. What this ease may conceal is that the characteristics of the topic and research question formulated at P1 were associated with downstream differences in the inferential patterns observed at P4: causal diction or overscoped phenomena at P1 tended to co-occur with the M3 patterns that became visible three stages later, while descriptively framed, appropriately bounded P1 questions tended not to. A causal question formulated at P1 plausibly creates pressure at every later stage to produce causal evidence, even when the design can only sustain descriptive observation, though the Cycle 1 artefact data can establish this co-occurrence but not the underlying causal sequence. This pattern is observational rather than experimentally established: the current design does not rule out alternative explanations such as differences in students' prior research-methods exposure, time pressure at P1 relative to later stages, or variation in statistical/quantitative literacy across groups any of which could independently affect both the sophistication of P1 framing and the M3 patterns downstream. Distinguishing an early-framing mechanism from these confounds is proposed as a specific design hypothesis for Cycle 2, not a confirmed

causal pathway. Three of the six sampled groups illustrate the pattern directly: the AC anchor case moved from causal framing at P1 to M3a at P4, receding only at the intervening operationalisation stage where a different code (M4) dominated; the TKE anchor case moved from overscoped framing (M1b) at P1 to a combined M3a+M3c pattern in its final report; and the DP additional case, whose P1 framing was descriptively bounded, remained M3-free through P4. These three trajectories are illustrative rather than exhaustive; a systematic P1-to-P4/P6 cross-tabulation across all six groups is left for Cycle 2, alongside the independent double-coding noted above.

This is consistent with Hammond and Gibbons' (2005) account of macro-scaffolding as structuring the conditions for learning before the moment of performance. The current P1 worksheet (LK 1) structures phenomenon identification procedurally but does not make its epistemic demands explicit: students are not supported to judge whether their chosen phenomenon is empirically grounded, answerable, and proportionate in scope. This P1 paradox is itself a Cycle 1 design finding, and it suggests where the Cycle 2 intervention focus could be shifted: from P4, where M3 becomes visible, toward P1, where the framing choices associated with it are made. The proposed revision, formalised as Design Principle 2 in Table 4, is a phenomenon-grounding protocol requiring students to specify the observable indicator, data source, and type of claim their data can support before proceeding to the research question.

**Operationalisation as an Epistemic Bottleneck (P2–P3)**

If P1 was the stage students found easiest, P2 and P3 were where the analytical demands of mini-research writing appeared particularly challenging. Together, the artefact-coding and questionnaire findings indicate that students could name a research variable but could not consistently translate it into non-overlapping, measurable indicators, a gap the Cycle 1 worksheets did not fully bridge: the indicator definition task in LK 2A did not explicitly feed into the item-writing task in LK 2B. The bottleneck is epistemological rather than merely procedural, stemming from variable operationalisation being unfamiliar or insufficiently foregrounded in students' prior disciplinary learning.

The proposed response, formalised as Design Principle 3 in Table 4, addresses both dimensions: procedurally, a mandatory variable–indicator–item traceability gate at LK 2, completed before data collection; conceptually, an in-class worked example tracing a theoretical construct through its indicators to specific items, with explicit discussion of why each link is necessary. This combination is consistent with evidence that self-regulated-learning-based writing instruction in Indonesian higher education is most effective when it pairs metacognitive strategy training with explicit modelling of the target cognitive operation (Anggraeni et al., 2025).

**Peer Review as a Feedback Architecture Problem**

The high-perceived-value/low-perceived-quality pattern in the peer-review data is compatible with a feedback-architecture explanation, though the available data do not establish this over a competence-based explanation, since student feedback competence was not directly measured. Applying Carless and Boud's (2018) account of feedback literacy (appreciating feedback, making judgements, managing affect, taking action), students appeared willing to appreciate and act on feedback, and the pattern may reflect limited judgement about what to evaluate in a peer's work, more consistent with a knowledge gap than a disposition problem, within the feedback-architecture reading. In Malecka et al.'s (2022) terms, the Cycle 1 P5 design provided the opportunity for feedback exchange but not the criteria, processing support, or enactment structure needed to realise it.

If this feedback-architecture interpretation is borne out, the Cycle 2 response should not expand the review criteria but focus them: a broad rubric risks reproducing the Cycle 1 problem, with students exhausting their attention on surface-level items before reaching substantive ones. A focused four-criterion rubric targeting phenomenon grounding, variable–indicator–item traceability, claim–analysis alignment, and conclusion scope accompanied by paired strong–weak examples, is proposed as Design Principle 4 in Table 4.

Cross-referencing the two data sources provides convergent evidence rather than proof of a single mechanism. The 74% of respondents who cited difficulty connecting findings to theory at P4 provide convergent evidence that P4 was experienced as an inferentially demanding stage, alongside the artefact-level M3 pattern, the moment at which the inferential leap from data to claim is most tempting and the scaffolding least adequate, consistent with M3 functioning as a recurrent cross-programme pattern within the Cycle 1 dataset rather than being confined to particular groups or topics a pattern to be corroborated with a larger sample in Cycle 2. A residual tension remains around worked examples: although 'Examples in LK helped my understanding' received the highest mean among the scaffolding items, 'add more examples' and 'clarify instructions' were the two most frequent open-ended suggestions (24% and 28% of respondents respectively), indicating that the current example supply does not yet sufficiently support independent transfer. This is formalised as Design Principle 5 in Table 4.

**Design Principles for Cycle 2 (Table 4)**

DBR is distinguished from other practitioner research by its commitment to producing usable design knowledge: transferable principles about what designs address what learning problems, under what conditions (Hoadley & Campos, 2022). The Cycle 1 findings generate five such principles, summarised in Table 4 with their evidential basis and proposed Cycle 2 revisions.

**Table 4.** Cycle 1 Evidence, Interpretive Mechanisms, Design Principles, and Proposed Cycle 2 Revisions

| Cycle 1 evidence | Interpretive mechanism | Design principle | Proposed Cycle 2 revision |
|---|---|---|---|
| M3 appeared across all three programmes and multiple artefact-producing stages, including in structurally strong reports (12/26 artefacts). | Task completion does not ensure claim calibration: the scaffold supported report structure but did not make epistemic boundaries explicit. | Claim calibration should be made explicit at the point of interpretation rather than assumed to follow from structural sequencing. | Claim–analysis alignment checkpoint at LK 4: students specify the analysis type, observed pattern, and permitted verb range before drafting the discussion. |
| P1 was rated as the least demanding stage (M = 2.92), yet topic and question framing at P1 was associated with the inferential patterns observed at later stages. | Students may not have fully perceived the epistemic stakes of phenomenon selection and question framing. | Phenomenon framing should become an explicit epistemic challenge rather than a descriptive warm-up task. | Three-column verification matrix at LK 1: indicator, data source, and inferential scope; supported by contrasting examples of observable phenomena versus opinions and descriptive versus causal questions. |
| An operationalisation bottleneck appeared across P2–P3: M3b emerged at P2, M4 at P3, 35 respondents reported difficulty writing items, and 32 reported difficulty determining indicators. | Variable operationalisation may be unfamiliar or insufficiently foregrounded in students' prior disciplinary learning. | Variable–indicator–item traceability should be made visible and verified before data collection begins. | Traceability gate at LK 2: mandatory research question → variable → indicator → item mapping, supported by an in-class worked example and guided modelling. |
| Peer review was rated as useful, but comments were insufficiently specific (M = 3.54); field notes also showed that some students were still completing their own drafts during P5. | Without explicit criteria, modelling, and protected review time, students may default to surface-level comments or own-work completion. | Peer review should be criterion-referenced and focused on high-leverage reasoning moves rather than generic quality. | Focused four-criterion peer-review rubric: phenomenon grounding and question answerability; variable–indicator–item traceability; claim–analysis alignment; and conclusion scope. Include paired strong–weak examples for each criterion. |
| Worked examples were rated as the most helpful scaffolding element (M = 4.43), while students frequently requested additional examples and clearer instructions. | Existing examples support recognition but do not sufficiently support independent transfer; the fading sequence is too abrupt. | Worked examples should be followed by a structured fading sequence from modelled to independent construction. | Faded example sequence: complete example → partially completed example → guided completion task → independent construction, implemented across at least two stages. |

*Note. Design principles are context-sensitive propositions derived from Cycle 1 artefact coding, questionnaire data, and field notes in one D4 vocational polytechnic. They are proposed for testing and refinement in subsequent DBR cycles rather than as generalised prescriptions.*

Two of these five principles warrant a candid note on confidence. The verification matrix for LK 1 (Principle 2) is grounded in the clearest finding of Cycle 1: negative cases consistently featured more empirically bounded P1 framing than M3-positive cases; the mechanism is traceable and the expected direction of effect is clear. The claim–analysis checkpoint for LK 4 (Principle 1) rests on equally solid evidence, but its effectiveness depends on whether a structured verb-permission table produces genuine epistemological reflection or only surface compliance: a student who learns to select *'mengindikasikan*' over *'membuktikan'* without understanding why has not developed claim calibration, only a new form of stylistic mimicry. Cycle 2 should treat this checkpoint as necessary but provisional, and collect evidence designed to distinguish genuine calibration from procedural compliance, consistent with evidence that self-regulated writing instruction is most effective when it develops metacognitive judgement rather than only procedural checklists (Anggraeni et al., 2025).

Taken together, the five principles reposition the Indonesian Language course's scientific writing component: from producing a correctly formatted research report, toward reasoning scientifically about one's own findings, the underutilised potential that Putera and Shofiah (2021) identify and that this DBR programme is working to realise.

**Limitations**

Several limitations should be noted. The artefact corpus comprises 26 coded artefacts from six purposively selected focal groups and one additional comparative artefact at one D4 polytechnic; the findings and design principles are context-sensitive propositions rather than generalisable claims, pending replication elsewhere. The coded unit is the artefact, with artefacts nested within student groups; this group-based sampling preserves ecological validity but limits within-group granularity: M3's cross-programme distribution does not imply uniform within-group distribution. The questionnaire (n = 72) was administered at a single end-of-cycle time point and reflects perceptions rather than direct measures of reasoning quality, a risk illustrated by the peer-review finding itself: students experienced the process as meaningful even when the feedback received was not specific. The questionnaire findings should therefore be interpreted cautiously, as nonresponse bias cannot be ruled out. The stylistic-mimicry mechanism, identified through peer debriefing of a small number of cases, remains an interpretive hypothesis rather than a confirmed finding, and should be treated in Cycle 2 as a research question. The iterative codebook development described above carries a related confirmation-bias risk that the audit-trail instrument and peer debriefing mitigate but do not eliminate; independent double-coding of the full corpus is planned for Cycle 2 to address this directly. Finally, as a practitioner-researcher study, the findings combine privileged access to implementation context with inherent interpretive proximity; peer debriefing and a detailed audit trail were used to address this, though McKenney and Reeves (2018) note that this tension is a structural feature of DBR generally rather than a unique weakness of this study.

## CONCLUSION

This study reports Cycle 1 of a Design-Based Research programme investigating a six-stage scientific writing scaffold (P1–P6) implemented within the Indonesian Language course across three D4 vocational study programmes. The workflow proved feasible to implement across six sessions in all three programmes, absorbing one in-cycle adjustment without structural disruption, and enabled students to complete the full structural sequence of mini-research writing. Whether this translates into improved writing quality relative to prior, non-scaffolded practice is a separate question that this Cycle 1 design, without a baseline or comparison condition, cannot answer; it is the central empirical target of Cycle 2. The workflow did not, moreover, yet adequately scaffold three high-stakes epistemic moves, namely claim calibration, variable operationalisation, and substantive peer evaluation. This gap represents a design problem rather than a failure of student effort and therefore provides the primary warrant for Cycle 2 redesign.

Three contributions follow. Empirically, M3 (premature causal, correlational, and significance claims) emerges as a recurrent pattern within the Cycle 1 dataset, appearing in 46% of coded artefacts (a group-nested rather than individual-level figure; see Limitations); it was also observed in final reports across all three programmes. Conceptually, the P1 paradox refers to the possibility that the stage students find easiest may be the most analytically consequential. This is proposed as a design hypothesis for Cycle 2, redirecting the intervention focus from P4 toward P1. Design-wise, five context-sensitive design principles (Table 4) reposition the course's scientific writing component from producing a correctly formatted report toward developing calibrated scientific reasoning, an underutilised potential the next DBR cycle will test.

## REFERENCES

Anderson, T., & Shattuck, J. (2012). Design-based research: A decade of progress in education research? *Educational Researcher*, *41*(1), 16–25. https://doi.org/10.3102/0013189X11428813

Anggraeni, C. W., Mujiyanto, J., Rustipa, K., & Widhiyanto. (2025). Effects of utilizing self-regulated learning-based instruction on EFL students' academic writing skills: A mixed-method investigation. *Asian-Pacific Journal of Second and Foreign Language Education*, *10*, Article 15. https://doi.org/10.1186/s40862-024-00317-6

Barbara, S. W. Y., Afzaal, M., & Aldayel, H. S. (2024). A corpus-based comparison of linguistic markers of stance and genre in the academic writing of novice and advanced engineering learners. *Humanities and Social Sciences Communications*, *11*, Article 284. https://doi.org/10.1057/s41599-024-02757-4

Bassett, M., & Macnaught, L. (2025). Embedded approaches to academic literacy development: A systematic review of empirical research about impact. *Teaching in Higher Education*, *30*(5), 1065–1083. https://doi.org/10.1080/13562517.2024.2354280

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Doo, M. Y., Bonk, C. J., & Heo, H. (2020). A meta-analysis of scaffolding effects in online learning in higher education. *The International Review of Research in Open and Distributed Learning*, *21*(3), 60–80. https://doi.org/10.19173/irrodl.v21i3.4638

Hammond, J., & Gibbons, P. (2005). Putting scaffolding to work: The contribution of scaffolding in articulating ESL education. *Prospect*, *20*(1), 6–30.

Hidayat, R., Utomo, W. T., & Rahmawati, Y. (2023). Analisis kebutuhan mata kuliah Bahasa Indonesia program studi D-3 Perhotelan berdasarkan outcome-based education. *Jurnal Pendidikan Bahasa Dan Sastra Indonesia Undiksha*, *13*(4), 272–284. https://doi.org/10.23887/jpbsi.v13i4.69839

Hoadley, C., & Campos, F. C. (2022). Design-based research: What it is and why it matters to studying online learning. *Educational Psychologist*, *57*(3), 207–220. https://doi.org/10.1080/00461520.2022.2079128

Hyland, K. (2007). Genre pedagogy: Language, literacy and L2 writing instruction. *Journal of Second Language Writing*, *16*(3), 148–164. https://doi.org/10.1016/j.jslw.2007.07.005

Lingard, L. (2020). The academic hedge Part I: Modal tuning in your research writing. *Perspectives on Medical Education*, *9*(2), 107–110. https://doi.org/10.1007/s40037-019-00559-y

Malecka, B., Boud, D., & Carless, D. (2022). Eliciting, processing and enacting feedback: Mechanisms for embedding student feedback literacy within the curriculum. *Teaching in Higher Education*, *27*(7), 908–922. https://doi.org/10.1080/13562517.2020.1754784

McKenney, S., & Reeves, T. C. (2018). *Conducting educational design research* (2nd ed.). Routledge. https://doi.org/10.4324/9781315105642

Nurasman, M. S., Firmansyah, B., & Nasution, N. (2026). PjBL in academic writing instruction in higher education: Challenges, strategies, and synthesis of empirical findings. *Bahtera: Jurnal Pendidikan Bahasa Dan Sastra*, *25*(1), 51–61. https://doi.org/10.21009/bahtera.251.06

Putera, Z. F., & Shofiah, N. (2021). Model kurikulum kompetensi berpikir pada pembelajaran Bahasa Indonesia di perguruan tinggi vokasi. *Metalingua: Jurnal Pendidikan Bahasa Dan Sastra Indonesia*, *6*(1), 29–36. https://doi.org/10.21107/metalingua.v6i1.10094

Putera, Z. F., Suyono, S., Rofiuddin, A., Harsiati, T., & Shofiah, N. (2025). Critical writing patterns and learning strategies among vocational college students: Evidence from thesis writing. *Jurnal Pendidikan Vokasi*, *15*(2), 166–177. https://doi.org/10.21831/jpv.v15i2.76100

Republik Indonesia. (2012). *Undang-Undang Nomor 12 Tahun 2012 tentang Pendidikan Tinggi* [Law Number 12 of 2012 on Higher Education], Pasal 35 ayat (3) dan ayat (5). Lembaran Negara Republik Indonesia Tahun 2012 Nomor 158.

Stubley, R., & Papen, U. (2026). Writing in vocational disciplines: Values, practices and post-compulsory teacher education. *Journal of Vocational Education and Training*, *78*(4), 645–661. https://doi.org/10.1080/13636820.2025.2597333

Syahril, Purwantono, Wulansari, R. E., Nabawi, R. A., Safitri, D., & Kiong, T. T. (2022). The effectiveness of project-based learning on 4Cs skills of vocational students in higher education. *Journal of Technical Education and Training*, *14*(3), 29–37. https://doi.org/10.30880/jtet.2022.14.03.003

Taub, M., Banzon, A. M., Outerbridge, S., Walker, L. R., Olivera, L., Salas, M., & Schneier, J. (2023). Towards scaffolding self-regulated writing: Implications for developing writing interventions in first-year writing. *Metacognition and Learning*, *18*(3), 749–782. https://doi.org/10.1007/s11409-023-09357-8

Wienanda, W. K., Harnita Sari, Y. I., Liani, T. A., & Utami, A. P. (2021). Menulis untuk publikasi: "Punya ide, tapi sulit menuliskannya" — analisis kebutuhan mata kuliah academic writing. *Jurnal Vokasi Indonesia*, *9*(1), 13–21. https://doi.org/10.7454/jvi.v9i1.217

## ACKNOWLEDGEMENT

The authors thank the students of the Accounting, Energy Conservation Engineering, and Tourism Destination Management study programmes at Politeknik Negeri Bandung for their participation, and the independent Indonesian Language lecturer who contributed to the peer debriefing process during artefact analysis.

## AUTHOR CONTRIBUTION STATEMENT

MSN conceptualised the study, designed the P1–P6 scaffolding framework, led the artefact-based analysis, drafted the manuscript, and undertook all remaining aspects of the study not attributed to the co-authors below. SNY and H provided academic guidance, reviewed the research framework, and supported the refinement of the manuscript. TH

facilitated and coordinated the peer debriefing session during artefact analysis, which was conducted by an independent Indonesian Language lecturer who is not an author of this article. All authors read and approved the final manuscript.

## AI DISCLOSURE STATEMENT

The authors used generative AI tools during the preparation of this manuscript to support language refinement, improve academic writing style, enhance clarity and coherence, and discuss the organisation of the manuscript. Artificial intelligence was not used to generate, analyse, or interpret the research data, nor to make scientific decisions or draw conclusions. All study design, data analysis, interpretation of findings, and the final content of the manuscript were conducted and verified by the authors, who accept full responsibility for the accuracy and integrity of the work.

## SUPPLEMENTARY MATERIALS AVAILABILITY

The following materials supporting the findings of this study are available from the corresponding author upon reasonable request: the P1–P6 scaffolding worksheets (LK1–LK4 templates); the structured field-note template used across sessions; the full reflection questionnaire instrument; the eight-aspect analytic scoring rubric (Appendix A) with its 0–3 scoring scale; illustrative coded artefact excerpts; and the complete artefact-level coding matrix underlying Table 2.

***Muhammad Syaiful Nurasman (Corresponding Author)**
Humanities Studies Group, Department of Accounting,
Politeknik Negeri Bandung,
Jl. Gegerkalong Hilir, Kabupaten Bandung Barat, West Java, 40559, Indonesia
Email: m.syaiful.n@polban.ac.id

**Sri Nur Yuliyawati**
Humanities Studies Group, Department of Civil Engineering,
Politeknik Negeri Bandung,
Jl. Gegerkalong Hilir, Kabupaten Bandung Barat, West Java, 40559, Indonesia
Email: srinuryuli@polban.ac.id

**Hazma**
Humanities Studies Group, Department of Accounting,
Politeknik Negeri Bandung,
Jl. Gegerkalong Hilir, Kabupaten Bandung Barat, West Java, 40559, Indonesia
Email: hazma@polban.ac.id

**Tedy Heriyadi**
Humanities Studies Group, Department of Refrigeration and Air Conditioning,
Politeknik Negeri Bandung,
Jl. Gegerkalong Hilir, Kabupaten Bandung Barat, West Java, 40559, Indonesia
Email: tedy.heriyadi@polban.ac.id

**Appendix Article Codebook: Analytical Code Definitions, Assignment Criteria, and Boundary Notes**
This codebook defines the analytical codes applied in the artefact-based analysis reported above (see Table 2). Codes were developed iteratively through the coding process and finalised following peer debriefing. All codes applied in the formal coding table (Table 2) have a status of VERIFIED or FINAL in the Koding_Artefak_Article_v1 sheet of the research workbook. Codes are non-mutually exclusive; a single artefact may receive multiple codes simultaneously. Positive contextual variations documented during coding are reported in the narrative as contrasting cases and are not assigned formal codes in Table 2.

**Table A1.** Codebook: Definitions, Criteria, and Boundary Notes

| Code | Name & Operational Definition | Assignment Criteria, Boundary Notes & Example |
|---|---|---|
| **M1a** | ***Opinion-based phenomenon selection*** The research phenomenon or problem statement is grounded in subjective opinion, personal perception, or moral evaluation rather than a directly observable, empirically bounded condition. | (1) Phenomenon described in terms of personal belief or normative judgement; (2) no reference to observable data or a measurable indicator; (3) the research question cannot be answered by data collection, only by argumentation. Distinguish from M1b (overscoped but observable): M1a phenomena are not falsifiable by data; M1b phenomena are falsifiable but exceed the capacity of a mini-research design. Example: 'Mahasiswa sekarang kurang disiplin' (students nowadays lack discipline) — opinion, not an observable condition. |
| **M1b** | ***Overscoped phenomenon or research question*** The phenomenon or research question encompasses more than one distinct issue, object, or analytical unit simultaneously, or implies a scope beyond what a single mini-research project can address. | (1) Multiple phenomena combined in a single research question; (2) the question implies four or more sub-questions; (3) the phenomenon requires a specialised instrument or multi-site design beyond available resources. Distinguish from M1a (may be empirically valid but too broad) and from M3b (M1b is a framing problem at P1; M3b is a claim problem at P2–P4). Example: a P1 artefact whose phenomenon combined energy consumption of lights, solar light, and toilet water (three distinct systems in one study). |
| **M3a** | ***Premature causal claim*** The artefact states or implies a causal, directional, or effect relationship between variables using language such as 'memengaruhi', 'menyebabkan', 'meningkatkan', 'berpengaruh', or 'terbukti', when the design uses only descriptive data collection. | (1) Causal verb or framing present in the research question, objective, discussion, or conclusion; (2) descriptive design (survey, observation, frequency count); (3) no pre-/post-test, control group, or regression analysis used. Distinguish from M3b (A↔B correlation vs. M3a's A→B causation) and M3c (significance language, not directionality). Example: a P1 research question, 'Apakah kemudahan bertransaksi melalui e-wallet meningkatkan intensitas impulsive buying?' — causal framing for descriptive survey data. |
| **M3b** | ***Relational or correlational claim beyond design capacity*** The artefact states or implies a correlational, relational, or associative relationship between variables ('berhubungan', 'berkorelasi', 'ada hubungan antara') when the design does not include a correlation analysis or equivalent statistical procedure. | (1) Correlational language in the research question, objective, or conclusion; (2) perception-based or descriptive questionnaire without a planned correlation analysis; (3) no statistical association test (e.g., Pearson r, Spearman ρ) used or planned. Distinguish from M3a (weaker but still unsupported relational claim) and K (K is the absence of a link; M3b is a false claim of one). Legacy M3c maps to Article M3b. Example: a P2 research question, 'Apakah kurangnya fasilitas berkorelasi negatif terhadap pencapaian?' — correlational framing without a correlation design. |
| **M3c** | ***Statistical significance without inferential basis*** The artefact uses language of statistical significance or certainty — 'signifikan', 'terbukti', 'secara signifikan', 'bermakna secara statistik' — when the analysis uses only descriptive statistics. | (1) Significance language present in the discussion, abstract, or conclusion; (2) the analysis section reports only descriptive statistics (mean, frequency, percentage); (3) no inferential test (t-test, ANOVA, chi-square, regression) conducted. Distinguish from M3a (significance language, not causal directionality; the two may co-occur within a single artefact) and R (assign M3c+R when 'signifikan' functions as stylistic mimicry rather than inferential intent). Legacy M3d maps to Article M3c. Example: a P4 artefact in which 'pengaruh FOMO signifikan' was claimed from descriptive Likert means without an inferential test. |
| **M4** | ***Indicator–item misalignment*** The instrument items (questionnaire questions, observation checklist items) do not trace clearly back to the indicators defined for the research variable, or the indicator definitions are too abstract for items to be derived without a missing intermediate step. | (1) An item asks about construct X but the variable/indicator concerns construct Y; (2) indicator definition remains abstract, without an observable proxy; (3) the variable→indicator→item chain has a missing or disconnected link. Distinguish from M3b (M4 concerns instrument construction, not claim language; may co-occur) and K (K is the absence of a theoretical link; M4 is the failure to operationalise an existing link into measurable items). Example: a P3 artefact in which impulsive-buying indicators were defined abstractly, with items partially replicating the indicator definition rather than measuring it. |
| **K** | ***Conceptual gap or missing element*** A required conceptual or theoretical link in the research argument is absent: the artefact names a variable, phenomenon, or construct but does not provide an operational definition, a theoretical basis, or a connection to the research question. | (1) Variable named but not operationally defined; (2) research gap stated without supporting contextual data or theoretical basis; (3) theoretical framework cited but not connected to variables or indicators. Distinguish from M4 (K is the absence of a link; M4 is a misaligned link) and M3b (K is silence about a relationship; M3b is a false claim about one). Legacy M3b (gap without contextual evidence) maps to Article K when no relational/causal claim is made. Example: a P3 artefact in which |

| Code | Name & Operational Definition | Assignment Criteria, Boundary Notes & Example |
|---|---|---|
| | | 'kesiapan memberi informasi destinasi' was named as a variable but not operationalised into non-overlapping indicators. |
| **R** | ***Language or register deviation*** The artefact uses language or register inconsistent with the genre conventions of the scientific report being produced — either adopting inferential or technical register without the methodological apparatus that legitimises it (stylistic mimicry), or deviating in the opposite direction (uncleaned template text, colloquial phrasing in formal sections). | (1) Inferential or significance language used in a purely descriptive context; (2) template placeholder text not replaced in the final submission; (3) colloquial language in the abstract, method, or conclusion. R is assigned only when independently documentable from the artefact text; assign M3c+R when significance language also functions as a register deviation. Not assigned for minor stylistic variation. Example: a P6 final report in which 'signifikan' was used descriptively ('86% of respondents') alongside uncleaned template text — compound code M3c+R by peer-debriefing consensus. |

*Note on M4 vs. K boundary. K is assigned when the required conceptual link is absent, a variable is named but the indicator chain leading from it is not established. M4 is assigned when the chain exists but is misaligned, indicators are defined but items do not adequately measure them. In practice, K often appears at P2 (variable defined, indicators absent) and M4 at P3 (indicators defined, items misaligned); co-occurrence is possible, and reviewers assign whichever code best captures the primary problem, noting the secondary one interpretively.*
*Note on codes not appearing in Table 2. M3d (overgeneralisation beyond the sample) appears in the codebook but was not assigned to any artefact in the final verified coding; cases initially considered for M3d were reassigned to M3a+M3c upon direct artefact review. L (positive writing behaviour) was not assigned as a formal code in any artefact in the current corpus; positive contextual variations are reported in narrative only. M1a (opinion-based phenomenon selection) was likewise not assigned to any artefact in the final verified coding. K (conceptual gap) and R (register deviation) are retained in the codebook but are not tabulated in Table 2, which reports only the M-series codes central to the claim-calibration analysis; R is reported narratively where it co-occurs with M3c. The M-series numbering runs M1, M3, M4: no M2 category was retained in the final codebook.*