Available on: https://jsshrs.ep-journals.org/ E-ISSN: 3051-3561, P-ISSN: 3051-3553

Journal of Social Science and Human Research Studies

Volume : 02 Issue : 10 October-2026

Process Writing Meets AI: A Hybrid Model for Teaching German Academic Writing Structure in the Merdeka Belajar Era

Ambo Dalle
Faculty of Language and Literature, Universitas Negeri Makassar, Makassar, Indonesia

Johar Amir
Faculty of Language and Literature, Universitas Negeri Makassar, Makassar, Indonesia

Published Date: October 26, 2026 **Article DOI: 10.65150/EP-jsshrs/V2E10/2026-11**

ABSTRACT:

Purpose: This study examined whether a hybrid pedagogical model that combines process writing with generative artificial intelligence (AI) feedback improves German-as-a-Foreign-Language (DaF) learners' command of academic writing structure under Indonesia's Merdeka Belajar–Kampus Merdeka (MBKM) policy, in a model where the writer, not the AI tool, remains the decision-maker.

Methods: A mixed-methods, nonequivalent control-group quasi-experiment was conducted with 42 undergraduate German Language Education students in two intact classes (n = 21 each) over one semester. The experimental class followed a Hybrid Process-AI Writing Model (staged drafting, structured AI feedback, teacher conferencing); the comparison class received conventional process writing instruction. Essays were scored with an analytic rubric and analyzed with ANCOVA (pretest as covariate, Holm adjustment); interviews with eight focal students were analyzed thematically.

Findings: After controlling for pretest scores, the classes did not differ on the total structure score, $F(1, 39) = 0.70$, $p = .409$, partial $\eta^2 = .02$, or on any rubric dimension (all Holm-adjusted $p = 1.000$); effect sizes were small and inconsistent in direction (d from −0.33 to 0.18). Both classes improved substantially, and they were not equivalent at pretest on paragraph organization/argumentation.

Implications: The study found no evidence that the hybrid model outperformed teacher-only process writing within one semester; given its limited power, the null result is inconclusive rather than evidence of equivalence.

Originality: This is among the first comparison-group classroom studies of a hybrid process-AI model for DaF academic writing, reporting a transparent, pre-specified null result.

KEYWORDS: process writing, generative AI, German as a foreign language, academic writing structure, Merdeka Belajar

INTRODUCTION

Academic writing in German as a Foreign Language (Deutsch als Fremdsprache, DaF) programs demands that learners master not only grammatical accuracy but also the conventional three-part structure of German academic prose – *Einleitung* (introduction), *Hauptteil* (body), and *Schluss* (conclusion) – together with explicit thesis statements, cohesive paragraphing, and disciplinary argumentation. These conventions are not universal: contrastive rhetoric research has shown that academic texts in German and English differ in preferred organization and in how text structure is signalled (Clyne, 1987; Kaplan, 1966), and research on German academic text competence indicates that command of such conventions develops gradually through sustained, guided practice (Steinhoff, 2007). In Indonesian teacher-training institutions, German Language Education students often produce texts that are grammatically competent yet structurally underdeveloped, with vague theses, paragraphs lacking topic sentences, and conclusions that restate the introduction rather than synthesizing the argument. The pretest scores in this study, which averaged just under 60% of the maximum rubric score in both groups (Table 2), are broadly consistent with this concern, though a single pretest cannot establish it as a general pattern beyond this sample. Process writing pedagogy, which treats composing as a recursive cycle of planning, drafting, and revising rather than a single-draft product, has long been used to address such weaknesses, yet its effectiveness depends heavily on the quality and timeliness of the feedback students receive at each stage. The rapid spread of generative artificial intelligence (AI) writing assistants offers a new way of delivering fast, iterative feedback on structure and cohesion, but it also raises the question of how such tools should be positioned within, rather than in place of, established process pedagogy (Elturki, 2026). At the same time, Indonesia's Merdeka Belajar–Kampus Merdeka (Freedom to Learn–Independent Campus, MBKM) policy gives lecturers and study programs more room to redesign course delivery, creating both an opportunity and an obligation to design technology-integrated writing instruction that is pedagogically principled rather than adopted for its novelty. This study was undertaken in that context: it designs and empirically tests a hybrid process-AI writing model for teaching German academic writing structure.

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

The MBKM policy was launched through Ministerial Regulation No. 3 of 2020 on National Higher Education Standards (Kementerian Pendidikan dan Kebudayaan Republik Indonesia, 2020), which was later replaced by Regulation No. 53 of 2023 on higher education quality assurance (Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, 2023) while continuing to leave institutions room to organize flexible and innovative learning. Studies of the policy describe it as a shift toward more adaptable, student-centered higher education (Widodo et al., 2020; Yusriadi et al., 2024). This autonomy allows German Language Education programs to integrate generative AI formally into writing courses, but instructors need a tested instructional model rather than ad hoc AI use. Without such a model, AI risks becoming a shortcut that bypasses the cognitive work process writing is meant to cultivate (Elturki, 2026). A model that leaves the writer in charge of accepting, adapting, or rejecting AI suggestions is also consistent with self-determination theory, which holds that learning environments supporting learners' autonomy and competence foster more sustained engagement (Deci & Ryan, 2000). This study therefore designs, implements, and evaluates a Hybrid Process-AI Writing Model specifically for German academic writing structure.

This study addresses two research questions. RQ1: Does a Hybrid Process-AI Writing Model produce greater gains in German academic writing structure – overall and in thesis clarity, paragraph organization and argumentation, cohesion and coherence, and conclusion quality – than conventional process writing instruction without AI support? RQ2: How do students engage with AI feedback and teacher conferencing across the stages of the writing process? It was hypothesized that (H1) students receiving the hybrid model would show significantly greater improvement than the comparison group in the total structure score and in each of the four rubric dimensions, controlling for pretest scores, and that (H2) within a teacher-mediated process cycle, students would use AI feedback predominantly for structural revision rather than for idea generation.

LITERATURE REVIEW

Process Writing and Feedback in L2 Academic Writing

Process writing pedagogy rests on Flower and Hayes's (1981) cognitive process theory, which conceptualizes composing as an interaction of planning, translating, and reviewing processes that writers move through recursively rather than linearly. Zamel (1983) showed that advanced second-language (L2) writers compose in a similarly non-linear and exploratory way, and Hyland (2003) argued that L2 writers nevertheless need explicit, staged support for organizing ideas into culturally and disciplinarily appropriate text structures, because generic process instruction alone does not guarantee command of target-language rhetorical conventions. Feedback is central to this cycle. Hattie and Timperley's (2007) model distinguishes task-, process-, and self-regulation-level feedback and implies that effective instruction moves students from surface correction toward engagement with the processes and self-monitoring behind their texts. Whether learners actually benefit depends on their feedback literacy, that is, their capacity to understand, judge, and act on feedback (Carless & Boud, 2018).

Generative AI as a Source of Writing Feedback

Generative AI has recently been examined as a source of such feedback, with mixed and context-dependent results. In a quasi-experimental study with 48 university English learners, Escalante et al. (2023) found no difference in learning outcomes between ChatGPT feedback and human tutor feedback over six weeks and recommended a blended approach that draws on the strengths of both. Zhang et al. (2025) compared GenAI-only feedback with hybrid GenAI-plus-tutor feedback among 60 Chinese EFL students over 12 weeks: GenAI feedback improved grammar and sentence variety, but its effect on organization and critical thinking was limited, whereas the hybrid group made significant gains in organization, critical thinking, and sentence variety. Engeness and Gamlem (2025) approached AI-generated feedback from a cultural-historical perspective, treating it as a cultural tool and deriving design principles for AI environments that support students' development as learners. Elturki (2026) argues, at a conceptual level, for an AI-aware "Process Pedagogy 2.0" that keeps human decision-making, rhetorical reasoning, and visible thinking at the center of the writing process. Taken together, this literature suggests that AI feedback is most beneficial for L2 academic writing when it is embedded in staged, teacher-mediated process cycles rather than deployed as a stand-alone correction tool. Note, however, that this proposition has been tested mainly against AI-only conditions rather than against well-implemented teacher-only process instruction.

Research Gap

Despite this growing body of work, most empirical studies of hybrid AI-assisted writing feedback have been conducted in English-as-a-foreign-language contexts. Research on AI-supported writing in German is only emerging: Strobl et al. (2024), for example, examined how ChatGPT-generated model texts shaped the revision of summaries by 22 advanced L2 German writers at a Dutch university, but in a two-week classroom intervention rather than a semester-long instructional design. To the best of our knowledge, no study has yet tested a staged hybrid model for German academic writing structure – with its distinctive Einleitung–Hauptteil–Schluss convention, verb-second and subordinate-clause word order that interacts with sentence-level cohesion, and formal register expectations – in an Indonesian teacher-training program with a comparison group. This gap was checked, not assumed: a search of Google Scholar and PubMed/PMC-indexed sources, conducted in September 2026 with combinations of the terms ("German as a foreign language" OR "Deutsch als Fremdsprache") AND ("generative AI" OR ChatGPT OR "AI feedback") AND (writing OR Schreiben), returned quasi-experimental hybrid-feedback studies only in English-as-a-foreign-language contexts (e.g., Escalante et al., 2023; Zhang et al., 2025) and no comparison-group study of a staged hybrid model for German academic writing structure; a systematic review with a full search strategy and forward/backward citation tracking was not conducted, and a study outside the terms searched or the indexes checked could have been missed. Moreover, existing hybrid-feedback studies rarely connect their designs to national curriculum policy, even though policies such as MBKM directly shape the autonomy instructors have to redesign writing courses around new technologies.

MATERIALS AND METHODS

Research Design

This study used a mixed-methods, nonequivalent control-group pretest–posttest quasi-experimental design over one semester (14 meetings) in a German academic writing course. Two intact classes were assigned to an experimental group, taught with the Hybrid Process-AI Writing Model, and a comparison group, taught with conventional process writing instruction without AI support. Pre- and posttest essays were used to compare structural writing gains, and classroom observation and interviews were used to explain how students engaged with the model, following a

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

convergent parallel logic (Creswell & Plano Clark, 2018). Because random assignment was not possible, the quantitative strand follows the TREND reporting guidelines for nonrandomized evaluations (Des Jarlais et al., 2004).

Participants, Sampling, and Ethics

Participants were 42 third-semester undergraduate students in the German Language Education Study Program at a public teacher-training faculty (LPTK) in Indonesia, in two intact classes of 21 students each (experimental $n = 21$; comparison $n = 21$). All had completed at least two semesters of general German coursework and had received no formal instruction in German academic writing structure. The classes were pre-existing intact groups assigned to conditions by course scheduling rather than by individual random assignment, so the sample is a convenience sample of intact classes drawn from the third-semester cohort. Participation was voluntary, and written informed consent was obtained from all students. With 21 students per class, the design had adequate power (.80 at $\alpha = .05$) to detect only large between-group differences (unadjusted $d \approx 0.9$ or larger); it was not powered to detect small-to-moderate effects, a limitation addressed further below.

The Hybrid Process-AI Writing Model

In the experimental class, each writing cycle combined five stages (Table 1; Figure 1): planning, drafting, structured AI feedback on organization and cohesion, teacher conferencing on argumentation and disciplinary voice, and revision. Two design principles governed the model. The first was sequencing: AI feedback preceded teacher conferencing so that conferencing time could be devoted to argumentation, disciplinary voice, and register rather than to basic structural problems. The second was writer authority: students remained responsible for the final decision to accept, adapt, or reject each suggestion, with the teacher helping them judge suggestions against academic register and argument strength. The comparison class followed the same planning–drafting–revising cycle with teacher feedback only.

Table 1: Stages of the Hybrid Process-AI Writing Model and the Comparison Class Equivalent

Stage	Experimental class (Hybrid Process-AI Writing Model)	Comparison class (conventional process writing)
Planning	Student plans the text.	Same as experimental class.
Drafting	Student writes a draft.	Same as experimental class.
Structured AI feedback	Student obtains AI feedback on organization and cohesion.	Not provided.
Teacher conferencing	Teacher and student discuss argumentation, disciplinary voice, register, and which AI suggestions suit academic conventions.	Teacher feedback on the draft.
Revision	Student revises, deciding which AI and teacher suggestions to adopt.	Student revises on the basis of teacher feedback.

Note. **Cycles ran across meetings 3–11 and were recursive (see Figure 1).**

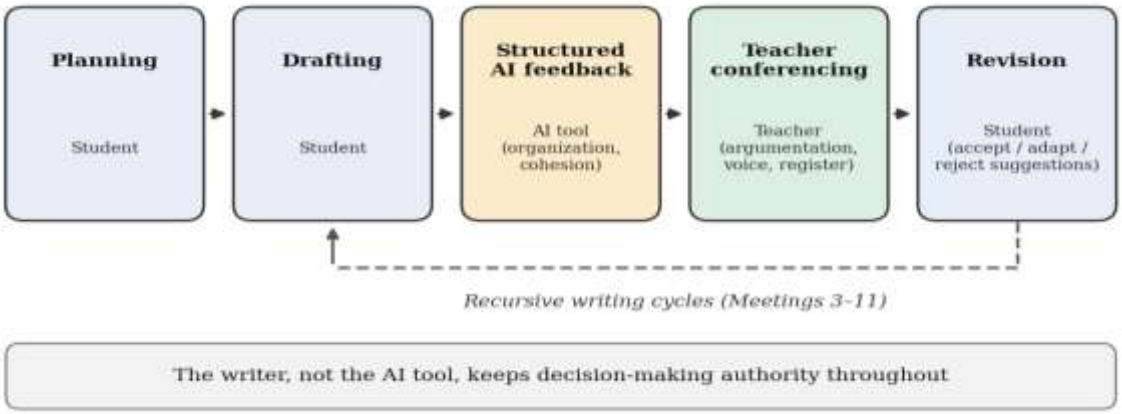


Figure 1: The Hybrid Process-AI Writing Model

Instruments

Writing structure was measured with an analytic rubric of four dimensions – thesis clarity (*Einleitung*), paragraph organization and argumentation (*Hauptteil*), conclusion quality (*Schluss*), and cohesion/coherence markers – each scored 0–25 for a maximum of 100 points. The rubric was developed from Hyland's (2003) genre-based writing criteria and validated by three DaF lecturers using Aiken's (1985) *V*, yielding a content-validity index of .86. A semi-structured interview guide, informed by Hattie and Timperley's (2007) feedback framework, explored how students used AI and teacher feedback across drafting stages, and a classroom observation checklist recorded the frequency and focus of teacher conferencing and AI-feedback sessions.

Procedures and Time Frame

The study ran for 14 meetings in one semester. Meetings 1–2 covered orientation and the pretest essay. In meetings 3–11, the experimental class followed the hybrid cycle described above, while the comparison class followed the same planning–drafting–revising cycle with teacher feedback only. In meetings 12–13, students produced a final academic essay under their condition-specific procedures, and meeting 14 comprised the posttest essay and semi-structured interviews with eight purposively selected focal students (four per group).

Data Analysis

Pretest equivalence was checked with an independent-samples *t* test. Posttest scores were analyzed with analysis of covariance (ANCOVA), with pretest score as covariate and group as the independent variable, for the total score and separately for each dimension; effect sizes are reported as

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

partial η^2 and Cohen's *d* (Cohen, 1988), with Holm's (1979) sequential procedure applied across the four dimension-level tests. All 42 students had complete pretest and posttest scores on all four dimensions, so no cases were excluded and no missing-data procedure was required. Diagnostic checks for each ANCOVA model (homogeneity of regression slopes, normality of residuals via the Shapiro–Wilk test, and homogeneity of residual variances via Levene's test) showed no indication of assumption violations (all $p > .05$; reported alongside Table 3). Interview transcripts and observation notes were analyzed with Braun and Clarke's (2006) six-phase thematic analysis (familiarization, initial coding, theme generation, review, definition, and reporting). The strands were integrated in a joint display during interpretation (Creswell & Plano Clark, 2018).

Scope and Limitations of the Methodology

Two intact classes at a single institution were assigned to conditions non-randomly, so condition is confounded with class (and possibly with instructor and meeting time), and students within a class are not independent observations. The design also compares hybrid instruction with teacher-only instruction, so the effect of AI feedback cannot be separated from the effect of receiving an additional feedback source. The study covered one semester and one generative AI tool. Together these features constrain causal inference and generalization to other institutions, languages, or AI platforms.

RESULTS

Baseline Equivalence

Pretest total scores did not differ significantly between the groups, $t(40) = 0.88$, $p = .382$ (experimental $M = 60.14$, $SD = 9.32$; comparison $M = 57.86$, $SD = 7.30$). At the dimension level, three of the four dimensions were equivalent at pretest (thesis clarity, $t(40) = 0.00$, $p = 1.000$; cohesion/coherence, $t(40) = 0.13$, $p = .896$; conclusion quality, $t(40) = -0.04$, $p = .967$), but the two classes were not equivalent on paragraph organization/argumentation, where the experimental class scored higher before instruction began, $t(40) = 2.12$, $p = .040$. This pretest imbalance is addressed as a limitation below and qualifies the interpretation of the organization/argumentation result.

Quantitative Results

The experimental class gained on average 17.4 points ($SD = 3.8$) in total structure score from pretest to posttest, compared with 18.7 points ($SD = 4.2$) in the comparison class (Table 2); the raw gain was therefore slightly larger in the comparison class. After controlling for pretest scores, the groups did not differ significantly on the posttest total score, $F(1, 39) = 0.70$, $p = .409$, partial $\eta^2 = .02$, adjusted difference (experimental − comparison) = −1.02 points, $d = -0.26$, 95% CI [−0.87, 0.35] (Table 3). None of the four dimension-level ANCOVAs reached significance either (thesis clarity: $F(1, 39) = 1.14$, $p = .291$; organization/argumentation: $F(1, 39) = 0.30$, $p = .589$; cohesion/coherence: $F(1, 39) = 0.78$, $p = .383$; conclusion quality: $F(1, 39) = 0.33$, $p = .570$), and all four remained non-significant after Holm adjustment (adjusted $p = 1.000$ in each case). Effect sizes were small and inconsistent in direction: negative (favoring the comparison class) for thesis clarity ($d = -0.33$), cohesion/coherence ($d = -0.27$), and conclusion quality ($d = -0.18$), and small and positive for organization/argumentation ($d = 0.18$) – the one dimension where the experimental class had already scored higher at pretest. Residual diagnostics gave no indication of assumption violations for any model (homogeneity of regression slopes, normality of residuals, and homogeneity of residual variances all $p > .05$). H1 was therefore not supported: the Hybrid Process-AI Writing Model did not produce significantly greater gains than conventional process writing instruction, overall or on any rubric dimension, in this sample.

Table 2: Descriptive Statistics and Mean Gains by Writing Dimension and Group

Writing dimension	Group	Pretest *M* (*SD*)	Posttest *M* (*SD*)	Mean gain (*SD*)
Thesis clarity (*Einleitung*)	Experimental	15.00 (3.56)	19.57 (3.87)	4.57 (1.63)
	Comparison	15.00 (3.35)	20.14 (3.57)	5.14 (1.80)
Organization/argumentation (*Hauptteil*)	Experimental	16.33 (3.09)	20.76 (3.24)	4.43 (1.86)
	Comparison	14.14 (3.60)	18.24 (4.44)	4.10 (2.23)
Cohesion/coherence markers	Experimental	14.76 (3.49)	18.57 (4.30)	3.81 (2.79)
	Comparison	14.62 (3.54)	19.10 (3.95)	4.48 (1.94)
Conclusion quality (*Schluss*)	Experimental	14.05 (3.81)	18.67 (3.86)	4.62 (2.33)
	Comparison	14.10 (3.51)	19.10 (3.49)	5.00 (2.24)
Total score (0–100)	Experimental	60.14 (9.32)	77.57 (9.44)	17.43 (3.78)
	Comparison	57.86 (7.30)	76.57 (7.10)	18.71 (4.20)

Note. M = mean; SD = standard deviation; n = 21 per group. Each dimension is scored 0–25. Mean gain = posttest M − pretest M. Pretest total scores were equivalent, t(40) = 0.88, p = .382; pretest scores on organization/argumentation were not, t(40) = 2.12, p = .040 (see text).

Table 3: ANCOVA Results by Writing Dimension (Pretest as Covariate)

Writing dimension	*F*(1, 39)	*p*	Holm-adjusted *p*	Partial η^2	Cohen's *d* [95% CI]
Thesis clarity (*Einleitung*)	1.14	.291	1.000	.03	−0.33 [−0.94, 0.28]
Organization/argumentation (*Hauptteil*)	0.30	.589	1.000	.01	0.18 [−0.43, 0.78]
Cohesion/coherence markers	0.78	.383	1.000	.02	−0.27 [−0.88, 0.34]
Conclusion quality (*Schluss*)	0.33	.570	1.000	.01	−0.18 [−0.78, 0.43]
Total score (0–100)	0.70	.409	—	.02	−0.26 [−0.87, 0.35]

Note. ANCOVA with pretest score as covariate and group (experimental vs. comparison) as the independent variable; positive d favors the experimental class. Holm-adjusted p values refer to the four dimension-level tests. None of the comparisons reached significance at α = .05. Diagnostic checks (homogeneity of regression slopes, normality and homogeneity of residuals) showed no indication of assumption violations for any model (all p > .05).

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

Qualitative Results

Thematic analysis of the interviews and observation notes generated three themes. *AI feedback as a structural mirror:* students described using the AI tool primarily to check whether their Einleitung stated a clear thesis and whether paragraphs in the Hauptteil followed logically from one another, rather than to generate new content. *Teacher conferencing as an anchor for judgment:* students consistently reported relying on their lecturer, not the AI, to decide which AI suggestions were appropriate for academic register and argument strength, particularly when suggestions conflicted with disciplinary conventions. *Evolving AI literacy:* several students described becoming more selective and critical users of AI feedback over the semester, moving from accepting most suggestions early on to evaluating suggestions against the rubric criteria in the later drafting cycles. Because only eight students were interviewed (four per group), the themes describe the experiences of a small purposive subset and cannot be generalized to the whole class.

Integration of the Strands

Table 4 compares the strands dimension by dimension. The pattern is one of divergence rather than convergence: students' accounts describe active, apparently helpful use of AI feedback for checking the thesis and paragraph logic, but the quantitative results show no corresponding between-group advantage on any dimension, including thesis clarity and organization/argumentation.

Table 4: Joint Display of Quantitative and Qualitative Findings

Rubric dimension	Quantitative result (experimental vs. comparison)	Qualitative evidence	Integration
Thesis clarity (*Einleitung*)	$d = -0.33$, ns; mean gain 4.6 vs. 5.1	Theme 1: students used AI to check whether the *Einleitung* stated a clear thesis	Divergent: perceived usefulness, no measurable advantage
Organization/argumentation (*Hauptteil*)	$d = 0.18$, ns; mean gain 4.4 vs. 4.1 (pretest not equivalent)	Theme 1: checking paragraph-to-paragraph logic; Theme 2: teacher judges argument strength	Inconclusive: only dimension with a same-direction (small) effect, but confounded by pretest imbalance
Cohesion/coherence markers	$d = -0.27$, ns; mean gain 3.8 vs. 4.5	No theme addresses cohesion markers directly; Theme 3 (selective use of suggestions) is indirectly relevant	Divergent / not addressed by qualitative data
Conclusion quality (*Schluss*)	$d = -0.18$, ns; mean gain 4.6 vs. 5.0	No dedicated theme in the current data	Unexplained

Note. Effect sizes and mean gains are taken from Tables 2 and 3. ns = not significant at $\alpha = .05$ (Holm-adjusted $p = 1.000$ for all four dimensions).

DISCUSSION

Contrary to H1, the experimental class did not outperform the comparison class on the total structure score or on any of the four rubric dimensions after controlling for pretest scores; on raw gain, the comparison class in fact improved slightly more on the total score. Both classes improved substantially from pretest to posttest (17–19 points on average out of 100), which indicates that the semester-long process writing cycle itself, common to both conditions, was associated with meaningful gains; adding structured AI feedback on top of that cycle was not associated with a detectable additional benefit in this sample.

This null result is compatible with Escalante et al. (2023), who found no difference in learning outcomes between ChatGPT feedback and human tutor feedback over six weeks and recommended a blended approach mainly on grounds of student preference rather than demonstrated superiority. It contrasts with Zhang et al. (2025), who found that a hybrid GenAI-plus-tutor condition produced significantly larger gains in organization and critical thinking than a GenAI-only condition. The comparison in the present study differs from both: the comparison class received teacher feedback without AI, not GenAI-only feedback, so a null result here does not contradict Zhang et al.'s finding that hybrid feedback beats AI-only feedback; it instead indicates that, in this sample, hybrid feedback did not clearly beat teacher-only feedback. One reading is that well-implemented teacher-mediated process writing, of the kind the comparison class received, may already capture much of the benefit that AI feedback is thought to add, at least for the structural dimensions measured here and within one semester.

Students' sense that AI feedback helped them check structure did not translate into significantly larger structural gains than in the comparison class, so the self-reports should be read cautiously alongside the quantitative result. The divergence between the strands shown in Table 4 admits two readings that cannot be distinguished with the present data: students may have perceived a benefit that the rubric did not detect (for example, in confidence or process rather than in the graded product), or perceived usefulness may not have translated into measurable gains within one semester and a single feedback tool. Revision-history or AI-conversation-log data, if retained, could help adjudicate between these two readings in a future analysis; this possibility is taken up under Suggestions for Future Research below.

Several features of the design limit how far this null result can be generalized or used to rule out a real effect. The study compared two single intact classes of 21 students, which is a modest sample for detecting anything short of a large effect (see Limitations); the classes were not equivalent at pretest on one dimension, organization/argumentation; and the qualitative theme that students used AI mainly to check structure did not correspond to a measurable structural advantage. Taken together, the most defensible conclusion is that this study did not find evidence for the hybrid model's superiority, not that the model has been shown to make no difference.

The qualitative finding that students relied on their lecturer, not the AI, to judge which suggestions fit academic register and argument strength is still consistent with Engeness and Gamlem's (2025) view of AI feedback as a cultural tool whose value depends on how it is embedded in a learning environment, and with Elturki's (2026) argument that AI-aware process pedagogy should keep human decision-making and rhetorical reasoning at the center of writing. However, because the quantitative results do not show an advantage for the class that used this AI-plus-teacher combination,

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

these theoretical accounts should be read here as describing how students engaged with the tools, not as an explanation of superior outcomes that this study did not find.

With respect to H2, the qualitative data were consistent with the expectation that students would use AI feedback predominantly for structural checking rather than for idea generation (Theme 1); however, this conclusion rests on the self-reports of eight focal students and was not verified against AI-conversation logs or revision histories.

Implications

On the evidence from this study, DaF instructors and program administrators should not adopt generative AI feedback on the expectation that it will outperform well-implemented teacher-only process writing instruction; the present sample provides no support for that expectation over one semester. Institutions considering AI-assisted writing feedback under the autonomy provided by MBKM would be better served by piloting it at small scale with pre-registered outcomes, as this protocol did, before committing curriculum resources, and by treating any reported gains from a single semester and a single class as provisional.

Research Contribution

This study contributes a transparently reported null finding to the hybrid AI-feedback literature, which has been concentrated in English L2 contexts and has more often reported positive effects (e.g., Zhang et al., 2025) than null ones (e.g., Escalante et al., 2023); reporting non-significant results, with the analysis code and data available for scrutiny, helps counter publication bias toward positive findings in this area. It also documents, for German as a Foreign Language academic writing specifically, a comparison of a theoretically motivated hybrid model (extending Elturki's, 2026, Process Pedagogy 2.0 framework) against teacher-only instruction under the institutional autonomy created by MBKM.

Limitations

This study relied on two intact, non-randomly assigned classes of 21 students each at a single institution, so condition is confounded with class and possibly instructor, and students within a class are treated as independent observations, which may overstate statistical precision; with one class per condition, the intraclass correlation cannot be estimated. The two classes were not equivalent at pretest on organization/argumentation, which qualifies that dimension's result specifically. With 21 students per group, the design had approximately 80% power to detect only a large unadjusted effect ($d \approx 0.9$) at $\alpha = .05$, and considerably less power for the four dimension-level tests after Holm correction; the observed effects ($|d| \leq 0.33$) are well below this threshold, so the null result mainly shows that if a true effect exists in this population, it is not large, rather than that no effect exists at all. The comparison was between hybrid and teacher-only instruction, so the specific contribution of AI feedback, as opposed to any additional feedback source, cannot be isolated. The one-semester duration does not permit conclusions about durability, and the use of a single generative AI tool means the findings may not generalize to other assistants. The interview data come from eight purposively selected students, self-reported perceptions are subject to social desirability bias, and no data on students' actual acceptance or rejection of AI suggestions were analyzed to corroborate the interview themes.

Suggestions for Future Research

Future research should replicate this comparison with more classes and greater statistical power, ideally with random assignment at the class or student level; include a non-AI comparison feedback source (e.g., structured peer feedback) and an AI-only condition, so that the contributions of AI and of teacher feedback can be separated; use raters blind to group and occasion; and add a delayed posttest to assess retention. Analyzing AI-conversation logs and revision histories directly, rather than relying only on self-report, would help explain the divergence observed here between students' perceived use of AI feedback and the absence of a measurable structural advantage.

CONCLUSION

This study examined whether a Hybrid Process-AI Writing Model would produce greater gains in German academic writing structure among Indonesian DaF learners than conventional process writing instruction. In this quasi-experimental comparison of two intact classes (21 students each), both classes improved substantially over the semester, but the hybrid class did not outperform the comparison class on the total structure score or on any of the four rubric dimensions after controlling for pretest scores, and the two classes were not equivalent at pretest on one dimension. Interview data suggested that students found AI feedback useful for checking structure, but this perception was not corroborated by a measurable advantage on the rubric. The study therefore does not provide evidence that adding structured AI feedback to a teacher-mediated process writing cycle improves German academic writing structure beyond what teacher-only instruction achieves, at least within one semester and with the sample size available here. Given the modest power of this comparison, the finding should be treated as inconclusive rather than as evidence that no effect exists, and it points to the need for larger, better-powered, and ideally randomized replications before generative AI feedback is adopted on efficacy grounds in DaF writing courses under the curricular autonomy of MBKM.

ACKNOWLEDGMENT

The author(s) would like to thank the German Language Education Study Program and the students who participated in this study for their time, openness, and cooperation throughout the semester. No external funding was received for this research.

ETHICAL AND COMPLIANCE STATEMENTS

Research ethics and informed consent: Participation was voluntary, and written informed consent was obtained from all participating students before data collection.

Conflict of interest: The author declares no conflict of interest.

Funding: This research received no external funding.

Plagiarism statement: This manuscript is original, has not been published elsewhere, and is not under consideration by another journal. Its similarity index, checked with turnitin, is 14% (≤ 15%).

Data availability: The anonymized score data and analysis code are available from the corresponding author upon reasonable request.

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

Corresponding Author: Ambo Dalle

Author contribution: Ambo Dalle (AD) conceptualized the study, designed the Hybrid Process-AI Writing Model, and drafted the manuscript. All authors reviewed and approved the final manuscript.

AI disclosure: The author(s) used Claude Sonnet 5 (Anthropic), in September 2026, during the preparation of this manuscript to assist with organizing the content according to the journal's template, drafting and language-editing sections of the text, running the pre-specified statistical analysis (independent-samples t tests and ANCOVA with Holm adjustment) on data supplied by the author, and checking the manuscript's internal consistency, based on the author(s)' study design and the essay scores the author collected and supplied; the AI tool did not generate, select, or alter any data values. After using the tool, the author(s) thoroughly reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

1) Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement, 45*(1), 131–142.
2) Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa
3) 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
4) Clyne, M. (1987). Cultural differences in the organization of academic texts: English and German. *Journal of Pragmatics, 11*(2), 211–247.
5) Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
6) Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.
7) Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11*(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01
8) Des Jarlais, D. C., Lyles, C., Crepaz, N., & the TREND Group. (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: The TREND statement. *American Journal of Public Health, 94*(3), 361–366. https://doi.org/10.2105/AJPH.94.3.361
9) Elturki, E. (2026). AI-aware process writing pedagogy: Rethinking the process-based approach. *TESOL Journal, 17*(2), Article e70127. https://doi.org/10.1002/tesj.70127
10) Engeness, I., & Gamlem, S. M. (2025). Exploring AI-driven feedback as a cultural tool: A cultural-historical perspective on design of AI environments to support students' writing process. *Integrative Psychological and Behavioral Science, 59*(1), Article 23. https://doi.org/10.1007/s12124-025-09894-8
11) Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education, 20*, Article 57. https://doi.org/10.1186/s41239-023-00425-2
12) Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365–387. https://doi.org/10.2307/356600
13) Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487
14) Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.
15) Hyland, K. (2003). *Second language writing*. Cambridge University Press.
16) Kaplan, R. B. (1966). Cultural thought patterns in inter-cultural education. *Language Learning, 16*(1–2), 1–20.
17) Kementerian Pendidikan dan Kebudayaan Republik Indonesia. (2020). *Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 3 Tahun 2020 tentang Standar Nasional Pendidikan Tinggi* [Regulation of the Minister of Education and Culture of the Republic of Indonesia Number 3 of 2020 on National Standards for Higher Education]. Kemendikbud.
18) Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia. (2023). *Peraturan Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia Nomor 53 Tahun 2023 tentang Penjaminan Mutu Pendidikan Tinggi* [Regulation of the Minister of Education, Culture, Research, and Technology of the Republic of Indonesia Number 53 of 2023 on Quality Assurance in Higher Education]. Kemendikbudristek.
19) Steinhoff, T. (2007). *Wissenschaftliche Textkompetenz: Sprachgebrauch und Schreibentwicklung in wissenschaftlichen Texten von Studenten und Experten*. Max Niemeyer Verlag.
20) Strobl, C., Menke-Bazhutkina, I., Abel, N., & Michel, M. (2024). Adopting ChatGPT as a writing buddy in the advanced L2 writing class. *Technology in Language Teaching & Learning, 6*(1), Article 1168. https://doi.org/10.29140/tltl.v6n1.1168
21) Widodo, S., Iqbal, M., Setiawan, B., & Handayani, L. (2020). The implementation of Merdeka Belajar Kampus Merdeka (MBKM) in Indonesian higher education. *Journal of Educational Science and Technology (EST), 6*(2), 130–138.
22) Yusriadi, Y., Farida, U., & Misnawati, M. (2024). "Merdeka Belajar Kampus Merdeka" policies in Indonesia higher education institutions: New public policy reform. *Al-Tanzim: Jurnal Manajemen Pendidikan Islam, 8*(1), 344–360. https://doi.org/10.33650/al-tanzim.v8i1.7173
23) Zamel, V. (1983). The composing processes of advanced ESL students: Six case studies. *TESOL Quarterly, 17*(2), 165–187. https://doi.org/10.2307/3586647
24) Zhang, Z., Aubrey, S., Huang, X., & Chiu, T. K. F. (2025). The role of generative AI and hybrid feedback in improving L2 writing skills: A comparative study. *Innovation in Language Learning and Teaching*. Advance online publication. https://doi.org/10.1080/17501229.2025.2503890

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

Corresponding Author: Ambo Dalle