INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN PROGRESSIVE EDUCATION & DEVELOPMENT

OPEN ACCESS JOURNAL

Vol 15, Issue 4, (2026) E-ISSN: 2226-6348

Learning Dissociation in AI-Mediated EFL Writing from a POA Perspective

Cao Yawen[1*], Wong Siew Ping[2], Nooreen Noordin[3]

[1]Faculty of Arts, Communication & Education, Kuala Lumpur University of Science and Technology (KLUST), Kuala Lumpur, Malaysia, [2]Department of Arts and Education, Faculty of Business and Human Sciences (FBiHS), Kuala Lumpur University of Science and Technology (KLUST), Kuala Lumpur, Malaysia, [3]Department of Language and Humanities Education, Faculty of Educational Studies, Universiti Putra Malaysia, Serdang, Malaysia

* Corresponding Author Email: caoyawen1206@foxmail.com

DOI Link: http://dx.doi.org/10.6007/IJARPED/v15-i4/29373

Published Online: 19 November 2026

Abstract

This study asks whether human-AI interaction on writing platforms can support the deep revision and selective learning that the Production-Oriented Approach (POA) sets as its goal. Working from the POA framework of assessment-as-learning, the study analysed the revision histories of 47 non-English-major undergraduates who used Pigaiwang, an AI writing evaluation platform. Revision depth was assessed from observable textual differences between successive drafts alone, without any inference about the AI feedback students received, and coded as Elaboration, Adoption, or Inertia. Two raters coded independently and reached high agreement (Cohen's $\kappa = .93$, with 95.7% exact agreement). In total, 87.2% of students produced no substantive elaboration, and the median score improvement was 0.50 points even though the overall gain was statistically significant, $t(45) = -3.89$, $p < .001$, Cohen's $d = 0.57$. Initial use of POA target words correlated negatively with score improvement ($r = -.294$). Learning type showed no significant association with initial proficiency, so revision engagement appears independent of language ability. The study contributes a codable framework for selective learning, names revision-learning dissociation as the gap between behavioural revision and cognitive learning, and proposes an optimised Teacher-Student Collaborative Assessment pathway for embedding AI feedback in POA-based writing instruction.

Keywords: Production-Oriented Approach, Assessment-as-Learning, AI Writing Feedback, Revision Behaviour, EFL Writing

Introduction

Research Background

Automated Writing Evaluation (AWE) systems have become a common supplementary tool in college English writing instruction in China. AI-powered writing platforms, Pigaiwang (Pigai.org) among them, are now used both in class and for after-class practice because they return instant feedback, sentence-level comments, and automated scores. What these platforms do best is spot surface-level language errors such as spelling mistakes, grammatical inaccuracies, and inappropriate collocations, and suggest a correction, which partly offsets

the shortage of teacher feedback in large classes. The same convenience brings a risk. Students often revise by accepting each AI suggestion as it stands and treating the automated feedback as an authoritative instruction. Their revision then stops at clicking to accept, with little reflection on the feedback and little engagement with deeper learning.

The picture looks different when set against the principles of the Production-Oriented Approach (POA), a foreign language teaching theory developed in China by Wen (2015). POA is built on a three-stage teaching cycle of motivating, enabling, and assessing, and it treats learning and use as one process. Assessment, in this view, is itself a phase of learning rather than the point at which instruction stops. Within POA's Teacher-Student Collaborative Assessment (TSCA), assessment carries an assessment-as-learning function: by evaluating and revising, students return to the language items taught in the enabling stage, reprocess them, and internalise them, so that input turns into output. Once the assessment stage is handed over to an AI platform, a question follows. Without teacher intervention, whether human-AI interaction on its own moves students toward the selective learning and deep processing that POA describes is what this study sets out to examine.

Problem Statement

Research on the learning benefits of AI writing feedback is divided. Several empirical studies report that AWE systems improve students' writing accuracy and overall quality. Others argue that telling feedback, which names the error and supplies the correct form, can suppress cognitive engagement and leave students accepting suggestions without thinking them through. What has received less attention is the question that matters for POA: how students actually respond to AI feedback, and whether their revision behaviour shows any of the assessment-as-learning that the approach advocates. Most studies ask only whether AI platforms raise scores.

Revision in L2 Writing

Revision has long been treated as a subprocess of writing rather than a separate stage (Flower & Hayes, 1981). Faigley and Witte (1981) proposed an influential taxonomy that divides revision into surface changes and meaning changes. Surface changes adjust form, for instance spelling, punctuation, and grammar, without altering the meaning of the text; meaning changes alter content, whether by adding or deleting material or by restructuring an argument. Later work has relied on this taxonomy. Min (2006), for example, operationalised revision as adoption, rejection, and ignoring, and then distinguished quality levels within each type.

Several factors shape the quality of revision in L2 writing. Barkaoui (2019) analysed L2 writers' pausing behaviour and found that revision strategies track the depth of cognitive processing: stronger writers pause longer and plan more during revision. Chen (2025) reviewed intervention studies on revision in writing classrooms and reported that explicit strategy instruction raises revision quality and writing performance, and helps students move from surface to meaning revision. Learner factors matter as well. Proficiency correlates with revision depth, since stronger writers attempt global and structural revisions while weaker writers stay with local, surface-level adjustments. Feedback factors also matter: the source of feedback (teacher, peer, or machine), its type (direct correction, indirect prompting, or metalinguistic explanation), and how focused it is all shape engagement. Research so far has

examined revision mainly under teacher and peer feedback, and has given little attention to how AI feedback, as a newer source, affects revision processes and cognitive engagement.

AI Writing Feedback and AWE Systems

Systems such as Criterion, Grammarly, and China's Pigaiwang now return instant feedback on spelling, grammar, vocabulary, sentence structure, and in some cases textual coherence. Their common design is telling feedback: the system locates an error and supplies a correction. Speed and coverage are what make this design useful.

The learning benefits of AI feedback remain contested. Parker et al. (2023) examined ChatGPT in academic writing assessment and found that although the tool returns language feedback at once, what determines learning is whether students can evaluate that feedback critically. Link et al. (2022) found that AWE systems raise revision frequency and writing quality, but the gain is confined largely to the accuracy of linguistic forms, with little change in content depth or argumentation. Ranalli (2021) added that the telling tendency of AI feedback encourages over-reliance, lowers cognitive engagement, and produces surface revision. Students who treat the system as an authoritative judge rather than a partner in dialogue tend to accept suggestions mechanically, and their revisions do not become deep learning.

AI feedback also carries a structural learning blind spot. Zhang and Hyland (2018) found that students attend to content and structure when the feedback comes from a teacher, but focus almost entirely on linguistic form when it comes from a machine. Systems built on natural language processing cannot judge whether the content of an essay responds to the objectives the teacher set. Within POA this blind spot matters more: once the assessment stage passes to an AI platform, the items that selective learning was meant to target may fall outside the scope of assessment altogether, and enabling and assessing come apart.

The POA Framework: Selective Learning and Assessment-as-Learning

The Production-Oriented Approach was developed over more than a decade by Wen Qiufang and her research team as a response to a long-standing problem in Chinese foreign language education, the separation between learning and use (Wen, 2015). Its instructional process runs through three stages. In motivating, communicative scenarios are designed to arouse students' learning needs and make their gaps visible. In enabling, the teacher breaks the production task into sub-tasks and supplies the vocabulary, target sentence patterns, and discourse structures needed to complete it. In assessing, TSCA is used to check whether students have actually used the enabled content in their production, and to deepen their learning in the process.

Two concepts in POA bear directly on this study: the selective learning hypothesis and the principle of assessment-as-learning. Selective learning holds that, faced with a large volume of input, learners must direct their limited attention to the small number of items that the current production task most requires (Wen, 2015). The enabling stage is therefore not designed for exhaustive coverage but around a carefully chosen set of high-value target items. Assessment-as-learning gives the assessment stage a function that the older checking role did not have: assessment is a learning activity in its own right. Revising their own work or that of their peers requires students to recall, identify, evaluate, and deploy the language taught in

the enabling stage, and that process internalises and consolidates what they have learned (Wen, 2018).

A full TSCA cycle has three phases. Before class the teacher selects typical samples for classroom assessment; in class the teacher leads focused discussion to identify problems and agree on revision plans; after class students carry out self-revision on the basis of that discussion. In practice the after-class phase often runs unguided: students revise on their own on AI platforms, without scaffolding or direction. That is where this study begins. If the after-class phase of TSCA is handed to an AI platform and left to run by itself, can assessment-as-learning still take place?

Self-Determination Theory and Revision Motivation

Revision depth also says something about why students engage with AI feedback. Self-determination theory (Deci & Ryan, 2000) places autonomous motivation, controlled motivation, and amotivation along a continuum of perceived self-determination, and the three revision-depth profiles sit on that continuum. Elaboration corresponds to autonomous engagement, where students expand their content in order to say more. Adoption corresponds to controlled adjustment driven from outside, in response to a score or a platform prompt. Inertia corresponds to amotivation, with minimal effort and a perfunctory submission. Read in motivational terms, assessment-as-learning depends on the quality of students' self-determined engagement during after-class revision.

Research Gaps and the Present Study

Taken together, the three strands of research leave three gaps. First, work on L2 revision has concentrated on teacher and peer feedback and has given little attention to revision shaped by interaction with AI. Second, work on AI feedback has concentrated on technical efficacy and has rarely asked how that feedback can be embedded in a complete instructional cycle, especially the alignment between content and objectives that AI cannot handle. Pan et al. (2025) proposed an AI-enhanced POA framework, but their work stays at the level of system design and reports no empirical test. Third, POA research has not yet examined what changes when AI enters the assessment stage: what happens to students' revision behaviour when the teacher is absent from after-class self-revision?

This study takes up those gaps by using selective learning and assessment-as-learning as analytical lenses and treating the revision histories students generated on an AI platform as evidence of human-AI interaction during the after-class phase of TSCA. An operational coding scheme turns the occurrence of selective learning into measurable indicators of revision behaviour, namely the distinction between revision that targets the enabled items and revision that does not. The design supplies the process perspective that AI writing feedback research has lacked, and gives POA an empirical basis for practice in AI-supported writing instruction.

Research Purpose and Significance

This study analyses the revision histories of 47 non-English-major undergraduates on Pigaiwang in order to describe the features of human-AI interactive revision, identify learning types, and propose an optimised TSCA pathway. Theoretically it tests selective learning and assessment-as-learning in AI-assisted writing and brings the phenomenon of revision-learning

dissociation into view. Practically it gives teachers a set of strategies for integrating AI feedback into TSCA.

Three questions guide the study. What revision behaviour and interaction patterns do students show on AI writing platforms? What learning types do those behaviours form, and how do they relate to language proficiency? What do the answers imply for the optimisation of TSCA?

Contributions of the Study

The study contributes to knowledge in three ways, two of them theoretical and one pedagogical. The first is conceptual. It identifies and names revision-learning dissociation, the condition in which students revise actively and raise their scores while their revisions carry no learning of the instructional targets. Revision has generally been read as evidence of engagement, on the assumption that more revision means more learning. The present data show that the two can separate. Students classified as Adoption revised repeatedly, improved their scores, and never touched the enabling targets their teacher had set. The dissociation is not a defect in the linguistic accuracy of the platform. It is a gap between what the platform is built to see and what the instruction intends to teach.

The second contribution is methodological. The study converts two POA principles, selective learning and assessment-as-learning, into codable indicators of revision behaviour. The three-type framework of Elaboration, Adoption and Inertia rests on two observable anchors, the change in text length and whether added content introduces new information. Because coding depends only on differences between successive drafts, it is independent of the content of the AI feedback and can be applied to revision histories exported from any platform. Selective learning has until now remained a design principle in POA rather than a measurable construct. Inter-rater agreement of $\kappa = .93$ across 47 records indicates that the framework is reliable enough to be reused.

The third contribution is pedagogical. The study specifies an optimised human-AI collaborative TSCA pathway and states the division of labour on which it rests. The platform checks linguistic form, the teacher selects cases that align with the enabling targets and leads critical discussion, and students keep a reflective revision log. The pathway responds to a specific finding, that assessment-as-learning did not occur when students revised alone, and it locates the missing element in the assessment stage rather than in student ability. Learning type showed no association with initial proficiency, which means the obstacle lies in instructional structure rather than in learner capacity.

These contributions differ from those of earlier work in three respects. Studies of AWE effectiveness (Link et al., 2022; Ranalli, 2021) ask whether scores rise, and generally find that they do. The present study accepts that result and asks what the rise contains. A significant mean gain of 5.66 points coexisted with a median gain of 0.50 and a negative correlation between target-word use and improvement ($r = -.294$), which shows that the aggregate figure conceals the distribution. Studies of L2 revision (Faigley & Witte, 1981; Min, 2006; Barkaoui, 2019) separate surface from meaning-level change but do not ask whether a meaning-level change serves an instructional target. The present framework adds that question. Work integrating AI into the POA cycle (Pan et al., 2025) proposes designs but does not examine

what students do with AI feedback when left to work alone. The present study supplies that evidence and shows why the design is necessary.

Methodology

Research Questions

The study addresses three research questions:

RQ1: What characterises the revision behaviour of non-English-major undergraduates on AI writing platforms, and what patterns of human-AI interaction emerge?

RQ2: On the basis of their revision behaviour and interaction patterns, into what learning types can students be classified, and how does the distribution of those types relate to their language production proficiency?

RQ3: What does the current state of human-AI interaction imply for the optimisation of POA's TSCA?

The questions are progressive. RQ1 describes phenomena, RQ2 moves to typological abstraction, and RQ3 turns to pedagogical application. Together they trace the path POA itself follows, from diagnosis at the assessment stage to the design of instructional intervention.

Research Context and Participants

The study took place in an English writing course for second-year non-English-major undergraduates at a private university in Guangdong Province. The course followed the POA model, and each unit was organised around an argumentative writing task taught through the motivating-enabling-assessing cycle. The unit examined here was themed Judging Truth and Falsehood, and students were asked to write an English argumentative essay of at least 200 words on whether a person's lying can be detected from appearance alone.

In the enabling stage the teacher explained and recommended eight vocabulary items (detect, deception, microexpression, mislead, nonverbal, perceive, suspicious, credible) and five target sentence patterns. The teacher told students that these items were resources rather than requirements, in keeping with the POA principle of selective learning: students decide whether and how to use the enabled language.

Participants were 47 second-year non-English-major undergraduates from two intact classes, aged 19 to 21. All had passed the College English Test Band 4 (CET-4) and could write a basic English argumentative essay. Students submitted their essays to Pigaiwang during the course and were free to revise them as often as they wished in response to platform feedback. At the end of the course the researcher exported all revision histories from Pigaiwang, which produced 47 records. Each student received an identifier (S01 to S47) for purposes of anonymity, and those identifiers are used throughout the analysis.

Data Collection

The data were the 47 revision histories exported from Pigaiwang. Each record held the full text of every version a student submitted, from the first draft to the final one, together with the platform's automated score for each version. Pigaiwang supplied the version history and the overall scores only; it did not export the sentence-level AI feedback that students had seen. Revision behaviour was therefore coded from the observable textual differences between successive drafts, with each pair of adjacent versions compared on the basis of what

the drafts actually contained. The comparison asked three things: whether the revision expanded the content and scope of the argument through new claims, evidence, examples, or fuller elaboration; whether it adjusted surface expression without changing the content; or whether it introduced no substantive change at all.

Processing ran in three steps. Revision traces from the 47 essays were extracted one by one into a structured analytical database, each record holding the student identifier, the text before revision, the text after revision, and the change between versions. Each revision history was then coded on its own using the revision-depth rubric set out below. For reliability, two coders, the researcher and a second rater who did not know the purpose of the study, coded all 47 histories independently against the same rubric. To keep the coding consistent, the researcher applied the criteria above throughout and re-checked ambiguous cases before deciding.

Data Analysis Framework

A single revision-depth framework was used to classify each student's dominant revision behaviour. Coding rests on two observable anchors: how much the text length changes, and whether new content adds information or restates what was already there.

Elaboration. The revision expands the content and scope of the argument by adding claims, evidence, examples, or fuller explanation, and the text usually grows substantially in length. This reflects deep processing: the student enriches the text beyond what the first draft contained.

Adoption. The revision adjusts expression, whether word choice, grammar, cohesion, or sentence structure, without altering the content or scope of the argument. Deleting a long faulty passage and rewriting it with the same meaning belongs here as well, because no new information is added. This reflects surface processing: the student responds to feedback without deepening the content.

Inertia. The revision brings no substantive change. Either the student submitted one version and never revised, or the differences between versions are limited to spacing, punctuation, or formatting. This reflects minimal processing or none.

Students who showed mixed behaviour were classified by their dominant behaviour across versions. Because the judgment rests only on observable differences between drafts, the scheme does not depend on the content of the AI feedback. Use of POA target words was treated separately as an outcome variable (see the correlation analysis below) and did not enter the revision-depth classification.

Data Analysis Procedures

Inter-rater reliability was established before the statistical analysis. Agreement reached 95.7%, or 45 of the 47 records, with Cohen's $\kappa = .93$ ($p < .001$), which counts as almost perfect agreement (Landis & Koch, 1977). Two borderline cases (S28, S47) were discussed and re-examined against the coding handbook, and the researcher's Inertia judgment was kept for both. All 47 records were coded, and all 47 students enter the descriptive accounts of revision behaviour. The platform published no score for S28's first draft, so that case has no initial score. It is therefore removed by listwise deletion from the four analyses that need one,

namely the paired t-test, the score-change comparison, the cross-tabulation of learning type with initial score level, and the correlations involving initial score or score change. Each of those rests on the same 46 cases. Table 2 reports the base for each variable separately. The coded data were then imported into SPSS 32.0 and analysed in six steps. Frequency analysis described the distribution of the three learning types. Descriptive statistics reported means, standard deviations, and medians for initial scores, final scores, score changes, number of versions, and number of revision acts. A paired-samples t-test compared students' initial and final scores in order to assess the overall learning effect of revision on the AI platform. An independent-samples t-test compared score improvement between the Elaboration and Adoption groups. Pearson correlation examined the relationships among initial scores, score changes, number of versions, number of revision acts, and the number of POA target words used in initial drafts. A chi-square test of independence examined whether learning type was associated with initial score level (Low, Medium, High).

Results

Overall Patterns of Revision Behaviour

Overall Distribution of Learning Types

A frequency analysis of the three learning types derived from the coding framework describes the general pattern of revision behaviour on the platform. Table 1 shows that Inertia was the largest group (n = 21, 44.7%), followed by Adoption (n = 20, 42.6%). Six students, or 12.8%, engaged in Elaboration.

Table 1

Frequency Distribution of Learning Types (N = 47)

Learning Type	Frequency	Percent (%)
Elaboration	6	12.8
Adoption	20	42.6
Inertia	21	44.7
Total	**47**	**100.0**

The distribution shows a clear pattern. Inertia and Adoption together account for 87.2% of students, none of whom elaborated their essays substantively, even though all of them had access to AI-generated feedback. Roughly one student in eight, 12.8%, produced the kind of deep, elaborative revision that POA treats as assessment-as-learning.

Descriptive Statistics of Revision Behaviours

Table 2 reports descriptive statistics for initial scores, final scores, score changes, and revision behaviour. The n column gives the base for each variable: initial score and score change were available for 46 students, because the platform published no score for S28's first draft, while final scores, version counts and revision counts were recorded for all 47.

The mean initial score was 70.50 (SD = 14.92, Mdn = 75.25) and the mean final score was 75.81 (SD = 14.87, Mdn = 82.00), an average improvement of 5.66 points (SD = 9.86). Score change varied widely (CV = 174.18%, skewness = 2.02, kurtosis = 3.06), and the maximum lay more than three standard deviations above the mean. The distribution of score changes was therefore highly skewed, and the median of 0.50 points describes the centre of that

distribution better than the mean does. A few students improved dramatically, while the typical student gained little.

For revision behaviour, students submitted a mean of 3.83 versions (SD = 4.67, Mdn = 2.00) and performed a mean of 7.34 revision acts (SD = 8.89, Mdn = 2.00). Both variables varied considerably, with coefficients of variation above 120%, which points to large individual differences in how much revision students did.

Table 2
Descriptive Statistics of Key Variables (N = 46-47)

Item	n	Min.	Max.	Mean	S.D.	Median
Initial Score	46	40.500	88.500	70.500	14.918	75.250
Final Score	47	40.500	92.000	75.809	14.872	82.000
Score Change	46	0.000	36.000	5.663	9.864	0.500
Number of Versions	47	1.000	24.000	3.830	4.673	2.000
Number of Revisions	47	0.000	29.000	7.340	8.894	2.000

Note. N = 46 for Initial Score and Score Change due to one missing value (S28).

A median score change of 0.50 points means that for more than half of the students AI-assisted revision produced little or no improvement, which matches the frequency analysis in which 87.2% of students made no substantive elaboration. How students engaged with the platform thus differed widely, and group comparisons are needed to say more.

Overall Effect of AI-Assisted Revision on Writing Scores

A paired-samples t-test compared initial and final scores in order to establish whether engagement with the AI writing platform improved writing scores. Table 3 reports the results for the 46 students with both scores.

Initial scores (M = 70.50, SD = 14.92) and final scores (M = 76.16, SD = 14.83) differed significantly, $t(45) = -3.89$, $p < .001$, 95% CI [-8.59, -2.73], and the mean improvement was 5.66 points. Cohen's d was 0.57 (Cohen, 1992), a medium effect by conventional benchmarks (0.20 = small, 0.50 = medium, 0.80 = large).

Table 3
Paired t-Test Results: Initial vs. Final Scores (n = 46)

Item	Initial Score	Final Score	Mean Difference	t	p
Initial Score - Final Score	70.50 ± 14.92	76.16 ± 14.83	-5.66	-3.894	< .001**

Note. Values are M ± SD. * p < .05. ** p < .01.

At the aggregate level, AI-assisted revision did improve essay scores significantly. The medium effect size (d = 0.57) shows that the improvement was moderate. Set beside the descriptive statistics reported above, in particular the median score change of 0.50 points, the result is less straightforward than the t-test alone suggests: the pre-post difference reached

significance mainly because a minority of students gained a great deal, while the typical student gained little.

This pattern matches the frequency distribution of revision-depth types reported above. The Elaboration students (n = 6, 12.8%) gained an average of 25.75 points and contributed disproportionately to the significant overall effect, while the Inertia students (n = 21, 44.7%), none of whom made a substantive revision, pulled the median down. A significant overall t-test therefore conceals wide variation in how much students benefited from AI feedback, and POA would expect that variation to follow from the quality of students' engagement with assessment-as-learning.

Differences Across Learning Types and Score Levels

Comparison of Score Change between Elaboration and Adoption Types

The Elaboration students improved by a mean of 25.75 points (SD = 11.82), the Adoption students by 5.30 points (SD = 5.69), a mean difference of 20.45 points. Levene's test showed that equal variances could not be assumed, so a corrected independent-samples t-test was used, and it found a significant difference, t(5.71) = 4.10, p = .007, Cohen's d = 2.76, 95% CI [8.5, 32.4].

For POA, this difference supports the proposition that the depth of engagement with assessment feedback determines learning outcomes: students who elaborated their content gained on average 20.45 points more than students who adjusted surface expression. The effect size (d = 2.76) shows how large that difference is in practice.

Table 4

Independent t-Test: Score Change by Learning Type

Item	Elaboration (n = 6)	Adoption (n = 20)	t	p
Score Change	25.75 ± 11.82	5.30 ± 5.69	4.10	.007**

Note. * p < .05. ** p < .01. Levene's test for equality of variances was significant (F = 14.15, p = .001). Therefore, the corrected t-value assuming unequal variances is reported.

Association between Learning Type and Initial Score Level

Table 5 cross-tabulates revision-depth type with initial score level for the 46 students who had an initial score on record. A descriptive pattern appears: Elaboration students clustered at the Low level (n = 4, 66.7%), with one student each at Medium and High. Adoption students were concentrated at Medium (n = 11, 55.0%), whereas Inertia students were distributed across all three levels (Low 35.0%, Medium 25.0%, High 40.0%). A chi-square test of independence showed no significant association between learning type and initial score level, $\chi^2(4) = 8.34$, p = .080. Because three cells had expected counts below five, a Fisher-Freeman-Halton exact test was also run and gave p = .093; neither test reached significance. Whether revision depth is related to initial proficiency therefore needs a larger sample to settle.

Table 5
Cross-Tabulation of Learning Type by Initial Score Level

Variable	Category	Elaboration (%)	Adoption (%)	Inertia (%)
Score Level	Low	4 (66.7)	3 (15.0)	7 (35.0)
	Medium	1 (16.7)	11 (55.0)	5 (25.0)
	High	1 (16.7)	6 (30.0)	8 (40.0)
Total		**6**	**20**	**20**

Note. Percentages are within each learning type. Score levels based on the 30th/70th percentiles of initial scores: Low < 65, Medium 65-81, High > 81. N = 46; the platform published no score for S28's first draft, so that case has no initial score level and is excluded here (see the data analysis procedures).

Relationships Among Language Proficiency, Revision Behaviour, and POA Target Word Usage
A Pearson correlation analysis examined the relationships among initial language proficiency, revision behaviour, and the use of POA target words. Table 6 reports the results.

Table 6
Pearson Correlation Matrix (N = 46-47)

Variable	Mean	S.D.	Initial Score	Score Change	Number of Versions	Number of Revisions	POA Words (Initial)
Initial Score	70.500	14.918	1				
Score Change	5.663	9.864	-0.339*	1			
Number of Versions	3.830	4.673	0.039	0.414**	1		
Number of Revisions	7.340	8.894	0.026	0.614**	0.653**	1	
POA Words (Initial)	1.383	1.295	0.323*	-0.294*	-0.147	-0.287	1

Note. * p < .05. ** p < .01.

The analysis produced four findings. Initial Score and Score Change correlated negatively (r = -.339, p = .021), which is consistent with the ceiling effect in writing assessment. Initial Score and POA Words correlated positively (r = .323, p = .028), so stronger students brought more target words into their first drafts without being prompted. Score Change and POA Words correlated negatively (r = -.294, p = .048): students who used more target words in the first draft gained less through revision. This is the evidence for revision-learning dissociation, since familiarity with target items at the outset did not carry through into effective revision or further learning. Revision quantity correlated positively with score change (r = .414 to .614, p < .01), but as the comparison across learning types shows, depth of revision predicted score gains far more strongly than frequency did.

Discussion

This study set out to examine, through POA theory, how non-English-major undergraduates revise on an AI writing platform, to identify distinct learning types, and to ask what conditions the occurrence of selective learning. The results reported a mixed picture of human-AI interaction. The platform did support score improvement overall, but that improvement was distributed very unevenly across student groups and was disconnected from the attainment of POA enabling targets. The discussion develops three arguments. It first proposes revision-learning dissociation as an account of the central findings. It then traces the origins of that phenomenon from three angles, the student, the platform, and the instruction. Finally it sets out an optimised pathway for embedding AI feedback in POA's Teacher-Student Collaborative Assessment.

Revision-Learning Dissociation: A New Challenge for POA in the AI Era

The central finding can be stated as one concept, revision-learning dissociation. Students respond to AI feedback and revise their texts at the behavioural level, but those actions are not matched at the cognitive level by selective attention to the enabling targets or by deep processing of them. Revision behaviour and target language learning therefore come apart. Quantitative and qualitative evidence both support the concept.

The most direct statistical expression of revision-learning dissociation is the significant negative correlation between the number of POA target words in initial drafts and the size of score improvement ($r = -.294$, $p = .048$). The more target vocabulary students used in the first draft, the less they gained through later revision. Had selective learning taken place in the assessment stage, target word use should have correlated positively with revision effectiveness: the more items students deployed while revising, the higher their gains should have been. The data run the other way. Target word use was already fixed at the drafting stage, and revision added no further depth or expansion. The median improvement of 0.50 points and the 87.2% of students who elaborated nothing point to the same conclusion. The significant overall result came mainly from large gains among a few Elaboration students; for most students the learning effect of AI-assisted revision was negligible.

The qualitative evidence points the same way. S01 submitted 24 versions, the most in the sample, and completed 25 revision acts, far above the mean of 7.34, moving from 57 to 83.5. None of those revisions touched a POA enabling target. They consisted of spelling corrections, tense adjustments, and simple lexical substitutions, which makes S01 a case of high investment with no deployment. S15, by contrast, submitted only 10 versions but deployed the target vocabulary throughout, using nonverbal more than 15 times, and added a content paragraph that took up the POA instructional objectives, gaining 32.5 points. The two cases show what revision-learning dissociation means: the amount of revision does not carry the quality of learning.

Revision-learning dissociation does not deny the technical efficacy of AI platforms. S01's case shows the opposite: 24 rounds of systematic revision turned an initial draft scoring 57 into a final version scoring 83.5, which is a real gain in the accuracy of linguistic forms. POA asks for more than correct forms, however. It asks that target language items transfer into production and be used creatively, and this is where AI platforms run into a structural limit. They help students revise correctly. They do not guide students toward learning deeply.

Why Assessment-as-Learning Does Not Occur Spontaneously

At the student level, revision is habitually equated with error correction, so students ask where the mistakes are and not which learning targets they have used. AI platforms reinforce this habit. Students attend to the formal corrections the system flags and overlook the vocabulary and sentence patterns taught in the enabling stage, since those items are never flagged as errors. Across all 24 versions of S01's revision, the eight target words and five target sentence patterns never entered the student's attentional scope. They were not too difficult; the structure of AI feedback simply never directed attention to them.

At the platform level, current AWE systems have a learning blind spot. Their feedback logic is built to identify linguistic deviations, and AI cannot judge whether an essay responds to the teacher's instructional objectives. It can flag a grammatical error, but it cannot tell a student that the argumentation has drifted from the issue under discussion in this unit, or suggest trying nonverbal and deception. The content dimension of assessment, which is what POA's selective learning centres on, therefore falls outside the reach of AI feedback. The telling form of that feedback also works against cognitive engagement. Once students grow used to spotting an error, clicking to accept, and gaining a higher score, the motive to think actively weakens. The dominance of compliance and the rarity of exploratory revision in this study follow from that tendency.

At the instructional level, the key structural factor is the absence of teacher guidance during assessment. Students revised on the AI platform on their own, without scaffolding. In POA terms this is a break in the TSCA cycle: the teacher taught the enabling targets carefully, then was absent from after-class self-revision. Students faced a system that could flag grammatical errors but could not understand instructional objectives, in place of a teacher who could direct their selective attention. The link between enabling and assessment therefore rested entirely on students' own metacognitive strategies, and the findings show that for most students this reliance did not hold.

From Revising for Scores to Revising for Learning: An Optimised TSCA Pathway

Teachers can position themselves as a bridge between AI feedback and POA instructional objectives, leaving the role of substitute or competitor aside. The optimised human-AI collaborative TSCA pathway below embeds AI feedback in the three phases of the POA assessment cycle.

Pre-class phase: the teacher reviews students' revision records and, guided by the unit's enabling targets, selects cases of instructional value. These might be revisions in which AI feedback aligns with an enabling target, misleading AI suggestions, or passages worth discussing that AI feedback did not address.

In-class phase: the teacher presents the selected cases and leads focused discussion. The question to pursue is not whether a sentence is correct but whether the AI's suggestion is reasonable, whether a better option exists, and which vocabulary and sentence patterns from the unit could be used there. The aim is to cultivate critical adoption, moving students from recipients of AI feedback to evaluators who keep ownership of their texts. With S01, for example, the teacher could show the grammar corrections and ask what the 26.5-point gain came from, whether any target vocabulary from the unit had been used, and where nonverbal

or deception might now be used. Questioning of this kind turns attention from whether the correction was right to whether the learning happened.

Post-class phase: students keep a reflective revision log, recording for each revision its type, its POA attribute, and the reason for accepting or rejecting the AI suggestion. Revision then becomes an explicit and reflective process rather than an automated one, and assessment gains a metacognitive dimension.

The logic behind this division of labour is that the platform checks linguistic form, which frees the teacher for work AI cannot do: selecting assessment materials that align with enabling targets, leading critical discussion, and designing reflective revision tasks. The arrangement uses the efficiency of AI while keeping assessment tied to the enabling targets. AI feedback on its own does not produce assessment-as-learning, but with teacher mediation, revision on an AI platform can serve that purpose.

Conclusion

Summary of Major Findings

Analysed through POA, the revision behaviour of 47 non-English-major undergraduates on Pigaiwang yields four findings.

First, the learning types separated clearly: 12.8% of students engaged in Elaboration, 42.6% in Adoption, and 44.7% in Inertia, so AI feedback did not by itself produce behaviour that promotes learning.

Second, the overall score improvement was statistically significant, yet the median gain was 0.50 points against a mean of 5.66, which shows that a minority of students carried the overall effect while most gained little.

Third, revision-learning dissociation appeared in the data: initial use of POA target words correlated negatively with score improvement ($r = -.294$). S01 illustrated it plainly, with 25 revision acts across 24 versions devoted entirely to linguistic form and none of them touching an enabling target, in spite of a 26.5-point gain.

Fourth, learning type showed no significant association with initial proficiency. Elaboration students appeared at both extremes of the score range, Adoption students clustered in the middle, and Inertia students were distributed across all three levels, which suggests that revision engagement does not depend on language ability.

Theoretical Contributions

The study makes two theoretical contributions. First, it tests and extends the POA concepts of selective learning and assessment-as-learning in the context of AI-assisted writing. Turning those concepts into codable indicators of revision behaviour, in the form of the Elaboration/Adoption/Inertia framework, gives empirical POA research a set of methodological tools that others can reuse. Second, it draws on the empirical data to propose revision-learning dissociation as a description of a pattern common among students on AI platforms, where revision is completed at the behavioural level without target language learning at the cognitive level. The concept accounts for the central findings here and offers a new analytical angle for future work on the learning effects of revision in AI-assisted writing.

Pedagogical Implications

Three implications follow for POA-based writing instruction. First, AI feedback belongs inside the POA assessment cycle. Because platforms check linguistic forms efficiently, teachers are relieved of sentence-by-sentence error correction and can spend their energy on what AI cannot do: selecting assessment materials aligned with enabling targets, guiding critical discussion, and designing reflective revision tasks.

Second, teachers need to cultivate students' capacity for critical adoption of AI feedback. In this study compliance-based revision dominated every interaction pattern, which indicates that students treat AI as an authoritative judge rather than a partner in dialogue. Within TSCA, teachers can guide students to question, compare, and select among AI suggestions, so that they become decision-makers in revision rather than executors of the technology.

Third, the assessment stage needs to value process as well as product instead of looking only at outcomes. The reflective revision log proposed here, in which students record the revision type, the POA attribute, and the reasons for accepting or rejecting an AI suggestion, is a concrete way to make that shift. Once revision behaviour itself becomes an object of instructional diagnosis and reflective learning, assessment on AI platforms can move from revising for scores to assessment-as-learning.

Limitations and Future Directions

The study has four limitations, each of which points to further research. First, the 47 participants came from two intact classes at a single university, and the Elaboration group held only six students, which limits statistical power and generalisability. Future work could use multi-institution designs with larger samples.

Second, revision depth was coded from observable draft-to-draft textual differences alone, which can show what students revised but not why. Future research could add think-aloud protocols, stimulated recall, or interviews to examine the cognitive and motivational processes behind the three revision-depth profiles.

Third, the design was naturalistic observation with no control group. The proposed human-AI collaborative TSCA pathway still needs testing in quasi-experimental studies that compare AI plus TSCA with AI-only self-revision.

Fourth, the study covered one writing task on one platform. Future research could use multi-task and multi-platform designs to examine how task type and feedback mechanism shape revision behaviour.

References

Barkaoui, K. (2019). What can L2 writers' pausing behavior tell us about their L2 writing processes? *Studies in Second Language Acquisition, 41*(3), 529-554.

Chen, J. (2025). *Writing and revising in second language classrooms: The role of self-regulation in cultivating expert writers*. Springer. https://doi.org/10.1007/978-981-96-1230-7

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

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.

Faigley, L., & Witte, S. (1981). Analyzing revision. *College Composition and Communication, 32*(4), 400-414.

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

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Link, S., Mehrzad, M., & Rahimi, M. (2022). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning, 35*(4), 605-634.

Min, H. T. (2006). The effects of trained peer review on EFL students' revision types and writing quality. *Journal of Second Language Writing, 15*(2), 118-141.

Pan, W., Mao, M., & Cheng, N. (2025). AI-enhanced production-oriented approach for feedback and assessment in EFL writing. In *Proceedings of the International Conference on Computer Science and Education* (pp. 74-87). Springer.

Parker, J. L., Becker, K., & Carroca, C. (2023). ChatGPT for automated writing evaluation in scholarly writing instruction. *Journal of Nursing Education, 62*(12), 721-727.

Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing, 52*, 100816.

Wen, Q. (2015). Constructing the theoretical system of the Production-Oriented Approach. *Foreign Language Teaching and Research*, 47(4), 547-558.

Wen, Q. (2018). The production-oriented approach to teaching university students English in China. *Language Teaching, 51*(4), 526-540.

Zhang, Z. V., & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing, 36*, 90-102.