Scientific Research Publishing

Open Journal of Social Sciences, 2026, 14(11), 155-176
https://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

The Transfer of Junior High School Students' English Writing in the Context of Generative AI Feedback and Its Relationship with Feedback Literacy and Self-Regulated Learning

Xiaoqing Sheng

College of Education, School of Teacher Education, Huzhou Normal University, Huzhou, China
Email: 18042208938@163.com

How to cite this paper: Sheng, X. Q. (2026). The Transfer of Junior High School Students' English Writing in the Context of Generative AI Feedback and Its Relationship with Feedback Literacy and Self-Regulated Learning. *Open Journal of Social Sciences, 14,* 155-176.
https://doi.org/10.4236/jss.2026.1411010

Received: October 23, 2026
Accepted: November 9, 2026
Published: November 12, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

To investigate the transfer performance of junior high school students' English writing under the condition of generative artificial intelligence feedback and its relationship with feedback literacy and self-regulated learning, 82 eighth-grade students were selected as the subjects, and the independent writing performance under the two teaching schemes of generative AI feedback and teacher feedback was compared using the unequal pre-test-post-test-delayed calibration experimental design of two natural classes, and the analysis was carried out in combination with questionnaires, feedback judgment tasks, and revised records. The results show that after controlling for pre-test writing scores, the adjusted average delayed-transfer score in the AI feedback class is 5.99 points higher than in the teacher feedback class ($P = 0.002$). There is a positive correlation between the change in feedback literacy and delayed transfer performance ($r = 0.413$, $P < 0.001$), and there is still a significant positive correlation after controlling the change of writing pre-test and self-regulation learning. The change in self-regulated learning is weakly positively correlated with transfer performance, but it does not reach statistical significance in the multiple regression model. The research shows that in the investigated classroom situation, students under the generative AI feedback scheme show high independent writing transfer performance, and feedback literacy may be an important correlation factor in understanding this phenomenon. Limited by the grouping of natural classes, the research results are not suitable for causal explanation.

Keywords

Generative Artificial Intelligence, English Writing, Writing Transfer, Feedback Literacy, Self-Regulated Learning, Junior High School Student

1. Introduction

The development of generative artificial intelligence (GenAI) provides new technical support for feedback in English writing instruction. Generative artificial intelligence based on large language models can provide suggestions on language error correction, content optimization, and text organization for learners' texts, and support writing modification through interaction. Kasneci et al. (2023) pointed out that large language models have potential applications in personalized learning support and teaching interaction, but their educational application still needs to pay attention to output reliability, learners' critical judgment, and teaching design. Guidance for generative AI in education and research issued by UNESCO also emphasizes that the people-oriented educational application principle should be adhered to, and learners' subjectivity and the appropriateness of technology use should be emphasized.

In English writing teaching, the educational value of generative artificial intelligence feedback depends not only on whether it can help students improve the current text, but also on whether the knowledge and strategies gained in the feedback process can be transformed into subsequent independent writing ability. According to the theory of learning transfer, the knowledge and skills acquired by learners in specific situations are not necessarily transferred to new task situations, and the occurrence of transfer is closely related to knowledge extraction, situational connection, and strategy application. Therefore, compared with the composition quality after artificial intelligence-assisted revision, the independent writing performance after removing tool support can reflect the learning effect of feedback activities from another angle.

This process also involves learners' active processing of feedback information. Carless and Boud (2018) proposed that feedback literacy is embodied in learners' ability to understand, judge, and use feedback information to improve their learning; Zimmerman's (2002) self-regulated learning theory emphasizes the role of goal setting, process monitoring, and self-reflection in learning activities. The two theoretical perspectives pay attention to the effective use of feedback information and the active adjustment of the learning process, respectively, which provides a theoretical basis for analyzing writing transfer under the condition of generative artificial intelligence feedback.

Based on this, this study takes junior two students as the subjects, adopts the unequal group pre-test-post-test-delay calibration experimental design of two natural classes, compares the changes in students' independent English writing performance and feedback processing behavior under the two teaching schemes of generative artificial intelligence feedback and teacher feedback, and investigates the relationship between feedback literacy, self-regulated learning, and writing transfer performance, to provide empirical reference for the rational application of generative artificial intelligence in junior high school English writing teaching.

2. Literature Review and Theoretical Basis

2.1. Generative Artificial Intelligence Feedback and English Writing Performance

Generative artificial intelligence (GenAI) provides a new technical path for second language writing feedback. Compared with the traditional automatic writing evaluation system, which mainly relies on preset rules to identify language errors, generative artificial intelligence based on large language models can generate explanatory suggestions according to writing tasks and text content, and support learners in carrying out revision through multiple rounds of interaction. However, the ability to generate feedback is not equal to its teaching effect, which still needs to be tested in combination with specific writing results and the learning process.

Parra Núñez and Castrillo de Larreta-Azelain (2025) systematically analyzed 30 empirical studies on ChatGPT English writing feedback published from 2023 to 2025, and found that they played a positive role in language accuracy, text organization, and overall writing quality. The systematic review of Sulistiyo et al. (2026) further points out that the effective use of generative artificial intelligence feedback is related to dialogue interaction, multi-manuscript revision, and clear evaluation criteria; Students are usually more likely to adopt partial language error correction, while deep revision involving content and structure requires extra judgment and teaching support. This shows that the teaching value of artificial intelligence feedback involves not only the feedback information itself, but also how learners understand and deal with feedback.

The comparative study of different feedback sources shows inconsistent results. Ekizoğlu and Demir (2025) conducted an eight-week study on 60 senior high school English learners in Türkiye and found that the artificial intelligence-assisted feedback group was better than the teacher feedback group in improving the total writing score, grammar, vocabulary, and coherence. However, another quasi-experimental study on 60 college English learners found that the writing scores of the artificial intelligence feedback group and the teacher feedback group both improved, but the difference between the two groups did not reach a statistically significant level. These results suggest that it is not appropriate to presuppose that generative artificial intelligence feedback is necessarily superior to teacher feedback, but to examine its actual performance in combination with teaching arrangements, learners' characteristics, and outcome indicators.

The measurement method of the feedback effect may affect the research conclusion. In the above study, pre-and post-writing tests were used to compare the teaching effect, but no delayed transfer test was conducted. Sulistiyo et al. (2026) also listed the long-term teaching effect as a problem to be further investigated. Therefore, in addition to immediate writing scores, it is helpful to further judge the learning effect associated with feedback activities by examining students' performance in completing new tasks after removing tool support.

2.2. English Writing Transfer and Its Measurement

Learning transfer refers to the process by which learners apply the knowledge, skills, or strategies acquired in a specific situation to other situations. Perkins and Salomon (1992) distinguish between near transfer and far transfer according to the similarity between learning situations and application situations, and emphasize that transfer may not only depend on the automatic application of familiar skills in similar situations, but also require learners to consciously extract and apply abstract principles.

In English writing teaching, original revision and independent writing transfer belong to different evaluation levels. Learners can improve the current text according to external feedback, which can reflect their completion of specific revision tasks; The independent use of relevant language knowledge and text organization strategies in new writing tasks can provide more direct evidence of transfer. Especially in the case that generative artificial intelligence can directly provide suggestions for revision, it is difficult to distinguish the direct contribution of external tools from the development of learners' own abilities if the learning effect is evaluated only according to the quality of the revised draft.

Therefore, this study defines English writing transfer as the performance of students applying the language knowledge and writing strategies obtained from previous feedback activities to new tasks without artificial intelligence and teachers' immediate help. The study examines the performance of independent writing through immediate post-test and delay test, and measures the recent transfer and its retention by writing activities with similar task requirements and different themes. Considering that there may be a correlation between task difficulty and writing performance, the test questions need to go through a difficulty assessment and scoring consistency test.

2.3. Feedback Literacy and Self-Regulated Learning

Feedback literacy foregrounds learners' active interpretation and use of feedback. Carless and Boud (2018) described four interrelated dimensions: appreciating feedback, making judgments, managing affect, and taking action. Their framework moves beyond a one-way transmission model and emphasizes learners' agency in constructing the meaning of feedback and using it to improve performance.

In GenAI-feedback contexts, students must evaluate whether a suggestion is accurate, appropriate to the task, and consistent with their intended meaning. The distinction between local and substantive revisions identified by Sulistiyo et al. (2026) further motivates examination of feedback judgment and revision behavior. This study therefore focused on feedback understanding, judgment, and effective revision. A reasoned rejection of an inappropriate suggestion was treated as evidence of judgment rather than automatically classified as a failure to take up feedback.

SRL theory provides a broader account of learner agency. Zimmerman (2002) described a cyclical process comprising forethought, performance, and self-reflec-

tion. In English writing, these phases can be understood in terms of goal setting and strategic planning, monitoring during drafting and revision, and evaluating outcomes to adjust subsequent strategies.

Feedback literacy and SRL are related but analytically distinct. The former concerns how learners interpret, evaluate, and act on external feedback; the latter concerns how they plan, monitor, and regulate their learning in pursuit of goals. In the present study, students' judgments and stated reasons for revisions informed the analysis of feedback literacy, whereas writing plans, monitoring practices, and reflective records informed the analysis of SRL. Both constructs may be associated with independent writing transfer, but the direction and strength of those associations require empirical examination and do not, by themselves, establish causal mediation.

2.4. Synthesis and Research Focus

Prior research provides evidence relevant to the use of GenAI feedback in English writing instruction, while indicating that findings vary across contexts and assessment measures. The quality of a revised text, immediate writing performance, and delayed independent writing performance capture different aspects of learning and should not be treated as interchangeable. The systematic literature reviews of Generative Artificial Intelligence Feedback have also shown closer attention to feedback uptake and longer-term outcomes.

From a theoretical perspective, feedback literacy concerns learners' understanding, judgement, and use of feedback, whereas SRL concerns goal setting, monitoring, and reflection during writing. monitoring, and reflection in the writing process. Combining the two is helpful for analyzing the performance of independent writing from two aspects: feedback processing behavior and learning process adjustment, although the empirical associations remain to be established.

Based on this, this study takes junior two students as the subjects and adopts the non-equivalent group pre-test-post-test-delay calibration experimental design of two natural classes to compare the changes in students' independent writing transfer and feedback processing behavior under the two teaching schemes of generative artificial intelligence feedback and teacher feedback, and to investigate the correlation between feedback literacy, self-regulated learning, and writing transfer. The main result of this study is the performance of independent writing without tools, combined with process evidence such as feedback judgment, composition revision, and learning reflection, to provide situational empirical reference for feedback teaching of English writing in junior middle schools.

3. Research Questions and Hypotheses

Drawing on the literature and theoretical framework above, the study addressed two research questions:

RQ1. How does independent English writing transfer performance differ be-

tween a GenAI-feedback class and a teacher-feedback class?

RQ2. How are feedback literacy and self-regulated learning associated with English writing transfer performance among lower secondary school students?

The following hypotheses were proposed:

H1. After adjustment for baseline writing performance, the GenAI-feedback class will outperform the teacher-feedback class on independent writing transfer.

H2. Higher levels of feedback literacy and self-regulated learning will be associated with better independent writing transfer performance.

4. Method

4.1. Participants and Research Design

A nonequivalent-group pretest-posttest-delayed-test quasi-experimental design was used. Participants were 82 Grade 8 students in two intact classes in City H. The experimental class (n = 41) received GenAI feedback, and the comparison class (n = 41) received teacher feedback. Both classes were taught by the same English teacher, used the same textbooks, and followed the same teaching schedule. Week 1 was devoted to the pretest, Weeks 2 - 7 to six weeks of writing-feedback instruction, Week 8 to the immediate posttest, and the delayed test to Weeks 3 - 4 after the immediate posttest.

To improve comparability, the study controlled for consistency among writing tasks, instructional content, class hours, revision opportunities, and assessment criteria. The study collected baseline writing scores and prior English achievement to describe initial differences. Because assignment was by intact class rather than randomization, class membership was confounded with feedback condition. Between-class comparisons are therefore exploratory analyses and do not support strict causal inference. Baseline writing scores, prior English achievement, and device access are presented in Table 1.

Table 1. Baseline writing performance, prior English achievement, and device access by class.

Measure	Indicator	GenAI feedback (n = 41)	Teacher feedback (n = 41)	t/χ^2	*P*
Writing performance	Baseline writing score	61.02 ± 12.27	60.62 ± 12.67	0.15	.884
English achievement	Semester final examination	96.20 ± 12.72	94.07 ± 10.57	0.82	.414
Device access	Personal learning device	33 (80.5%)	30 (73.2%)	0.27	.601
Device access	Home internet access	39 (95.1%)	37 (90.2%)	0.18	.672

Baseline writing was scored out of 100 by averaging two raters' assessments across task fulfillment, organization, vocabulary, grammatical accuracy, and appropriateness. Prior English achievement refers to the Grade 8 first-semester final examination (maximum 120). Continuous variables are presented as M ± SD and device-access variables as n (%). Independent-samples t tests and continuity-corrected chi-square tests were used. No statistically significant differences were detected on the baseline scores, English achievement, and devices, which does not establish equivalence on all learner characteristics or rule out unmeasured class differences.

4.2. Procedure

The study comprised four stages: pretesting, instructional intervention, immediate post testing, and delayed testing (Table 2).

Table 2. Study procedure.

Stage	Timing	Activities
Pretest	Week 1	Independent writing pretest; feedback literacy and SRL questionnaires
Intervention	Weeks 2 - 7	Three rounds of drafting, feedback, and independent revision
Immediate posttest	Week 8	Independent writing posttest; feedback literacy and SRL questionnaires
Delayed assessment	3 - 4 weeks after intervention	Independent writing transfer assessment without AI support

Both classes completed the same writing tasks following the sequence: independent drafting, feedback receipt, feedback evaluation, independent revision, and reflection.

The GenAI-feedback class used "seed 2.0 lite" to obtain writing feedback. A standardized prompt instructed the system: "You are a writing-feedback assistant for Grade 8 English learners. Based on four dimensions—task fulfillment, organization, vocabulary, and grammar—identify no more than three of the most important problems, explain why they are problematic, and offer suggestions that guide students to revise independently. Do not rewrite the entire text; acknowledge uncertainty when appropriate." Under teacher supervision, students read the feedback, recorded reasons for their judgments, and revised independently. De-identified prompts and outputs were retained.

In the control class, the English teacher provided written feedback using the same assessment dimensions. Students completed feedback judgments, revisions, and reflective records under the same requirements as the experimental class. Both classes were taught by the same teacher and used identical writing tasks, assessment dimensions, and revision procedures. To improve comparability, a common upper limit was set on feedback items per composition, and expectations for explanations and classroom revision time were standardized. The AI class received feedback through a standardized prompt; the teacher-feedback class received written comments in the same dimension from the teacher. Differences in actual wording and explanatory content may nevertheless have remained, so the feedback source and its implementation are considered jointly as components of each instructional approach.

4.3. Instruments

4.3.1. English Writing Transfer Assessments

Three independent writing assessments were administered: pretest (T0), immediate posttest (T1), and delayed transfer test (T2). No GenAI or other external feedback tools were permitted during any assessment in **Appendix A**.

Test tasks were designed according to Grade 8 English curriculum requirements and used the writing subject with comparable genres and different topics.

Writing was scored on task fulfillment, organization, vocabulary, grammatical accuracy, and appropriateness, for a maximum of 100 points.

Two experienced English-language educators independently rated the essays without knowing students' group membership. Interrater reliability was assessed using the intraclass correlation coefficient (ICC). Delayed-test performance was the primary outcome, while immediate posttest performance served as a supplementary outcome.

4.3.2. Feedback Literacy

The feedback literacy measure was informed by Carless and Boud's (2018) student feedback literacy framework and adapted to junior secondary English writing. It assessed feedback comprehension, judgment, and use through 20 items on a five-point Likert scale, from strongly disagree (1) to strongly agree (5). Higher scores indicate higher self-reported feedback literacy. Full items, dimension assignments, and scoring rules appear in **Appendix B**.

A situated feedback-judgment task complemented the questionnaire. Whereas the questionnaire assessed students' appreciation of feedback, confidence in evaluating suggestions, and tendency to use feedback, the situated task required students to assess proposed revisions and justify acceptance or rejection, thereby supplementing self-report measures.

4.3.3. Self-Regulated Learning

Drawing on Zimmerman's (2002) cyclical model, writing-related self-regulated learning was operationalized in three dimensions: goal planning, process monitoring, and self-reflection. The questionnaire comprised 15 five-point Likert items; higher scores indicated higher self-reported writing self-regulation. The complete instrument and scoring rules are provided in **Appendix C**.

Students' writing-reflection records collected during the intervention were also used as supplementary interpretive evidence.

4.3.4. Coding and Calculation of Process Indicators

Two process indicators were used. Because they differ in data source, denominator, and measurement occasion, their definitions are specified separately.

(1) Item-level effective revision rate. Data comprised all feedback items from three rounds of writing. Each student received three codable feedback items per round, yielding 369 items per class (41 students × 3 rounds × 3 items) and 738 items in total. Coding took place immediately after each revision round and concerned intervention-process data rather than test scores. Each item was coded for feedback accuracy, uptake, revision effectiveness, adequacy of justification, and justified rejection. A revision was effective only if it addressed the language or content issue identified, corrected or substantially improved the relevant passage, and introduced no new linguistic error or semantic distortion. Failure to meet any criterion meant that the revision was coded as ineffective. The effective revision rate was the number of effective revisions divided by all 369 feedback items in the class, regardless of uptake. If uptake items alone are used as the denominator, the

measure is instead the effectiveness rate among adopted items: 50.4% (127/252) in the AI-feedback class and 41.2% (94/228) in the teacher-feedback class; these figures are reported only as supplementary information. Two coders independently assessed 20% of items (148 items), with agreement $\kappa = 0.78$; disagreements were resolved by discussion.

(2) Effective revision proportion in the situated task. At each administration, students evaluated 12 suggestions concerning the same essay, decided whether to adopt each, justified their decisions, and implemented revisions for adopted suggestions. The task was administered at pretest (Week 1) and posttest (Week 8). The effective revision proportion was the number of effective revisions divided by 12, the fixed number of suggestions.

These indicators should not be compared directly. The first uses all classroom feedback items as its denominator and reflects overall revision output during instruction; the second uses standardized suggestions and reflects judgment and revision performance under controlled task conditions.

4.4. Data Collection and Analysis

Data included scores from the three independent writing assessments, feedback-literacy and SRL questionnaires, and writing revision records.

Baseline characteristics, writing scores across time, and questionnaire distributions were described, together with interrater and internal consistency reliability. An exploratory analysis of covariance (ANCOVA) then modeled delayed writing scores as the dependent variable and baseline writing scores as a covariate, reporting the adjusted between-class difference, effect size, and 95% confidence interval. Standard errors, significance tests, and confidence intervals from conventional ANCOVA assume, among other conditions, independent errors, whereas students within the same class may be correlated. Because each feedback condition involved only one intact class, treatment and class membership were perfectly confounded; class and feedback effects cannot be separated, nor can between-class sampling uncertainty be reliably estimated. Student-level *P*-values and confidence intervals are therefore descriptive of the fitted model and should not be interpreted as evidence of a generalizable intervention effect.

Correlations and parsimonious regressions examined associations of feedback literacy and self-regulated learning with independent writing transfer, with revision records used for interpretation. A sensitivity regression assessed whether the association between change in feedback literacy and delayed writing scores persisted after accounting for class differences and initial learner characteristics. In addition to baseline writing, changes in feedback literacy, and changes in self-regulated learning, the model included class membership (AI = 1; teacher feedback = 0), baseline feedback literacy, and baseline self-regulated learning. The stability and significance of the feedback literacy change coefficient were assessed. Given the sample size and intact-class design, no complex

multiple-mediation analysis was conducted and associations were not interpreted causally.

5. Results

5.1. Descriptive Statistics and Measurement Quality

The analyses included 82 Grade 8 students (41 per class). All students completed the pretest (T0), immediate posttest (T1), and delayed test (T2), with no missing observations. Writing scores, feedback-literacy scores, and SRL scores are summarized in Table 3.

Table 3. Descriptive statistics for the main variables by class.

Variable	GenAI feedback (n = 41)	Teacher feedback (n = 41)
Baseline writing (T0)	61.02 ± 12.27	60.62 ± 12.67
Immediate writing (T1)	71.54 ± 14.76	66.10 ± 16.26
Delayed transfer (T2)	71.15 ± 14.03	64.74 ± 16.34
Feedback literacy pretest	61.24 ± 10.94	60.90 ± 9.67
Feedback literacy posttest	73.76 ± 11.51	66.29 ± 12.78
SRL pretest	51.32 ± 7.12	50.78 ± 7.97
SRL posttest	59.49 ± 10.31	53.51 ± 11.65

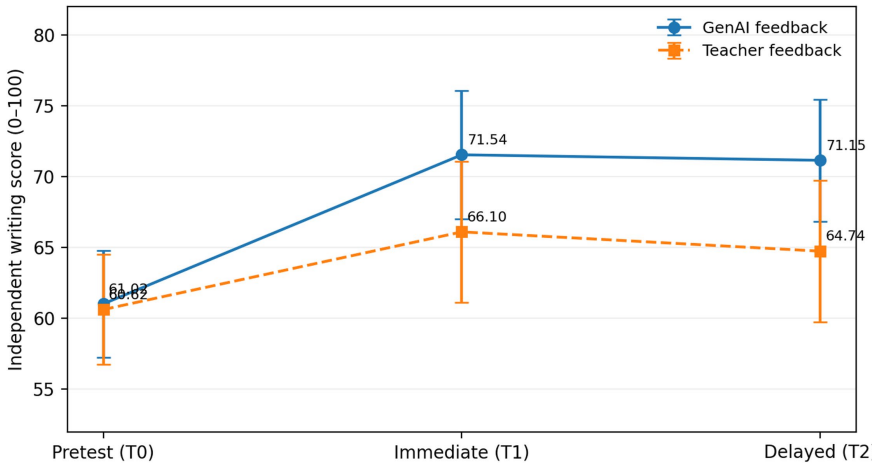
Writing assessments were scored from 0 to 100; the feedback-literacy questionnaire from 20 to 100; and the SRL questionnaire from 15 to 75. Values are M ± SD; n = 41 per class, with no missing data.

Interrater agreement for delayed-test writing scores was ICC = 0.93 (two-way random-effects model, absolute agreement), indicating acceptable reliability. Cronbach's alpha was 0.88 for the feedback-literacy questionnaire and 0.85 for the SRL questionnaire. Across the three dimensions of each questionnaire, alpha coefficients ranged from 0.76 to 0.81 and from 0.75 to 0.78, respectively. No items were identified as requiring removal or revision; the reported internal consistency supported subsequent analyses.

5.2. English Writing Transfer under Different Feedback Conditions

To address RQ1, writing performance was compared descriptively at baseline, immediately after the intervention, and at the delayed assessment.

In the GenAI-feedback class, the mean writing score increased from 61.02 (SD = 12.27) at T0 to 71.54 (SD = 14.76) at T1 and was 71.15 (SD = 14.03) at T2. Corresponding scores in the teacher-feedback class were 60.62 (SD = 12.67), 66.10 (SD = 16.26), and 64.74 (SD = 16.34). Both classes improved at T1, followed by small descriptive declines at T2 (0.39 and 1.35 points, respectively). Figure 1 depicts the trajectories.



Error bars represent 95% confidence intervals for the means. The vertical axis begins at 52 to make changes more visible.

Figure 1. Independent English writing scores across three assessments.

In an exploratory ANCOVA with T2 as the dependent variable and baseline writing score as a covariate, the estimated adjusted between-class difference was 5.99 points, 95% CI [2.31, 9.67], F(1, 79) = 10.49, $P = .002$, partial $\eta^2 = 0.117$. The unadjusted difference was 6.40 points (d = 0.42, $P = 0.061$). A supplementary ANCOVA for T1 yielded an adjusted difference of 5.03 points, 95% CI [1.16, 8.89], F(1, 79) = 6.68, $P = 0.012$, partial $\eta^2 = 0.078$ (d = 0.35). The reported correlation between baseline and delayed writing scores was r = 0.82; adjustment for baseline increased statistical precision relative to the unadjusted comparison.

Table 4. Exploratory Comparisons of Independent Writing Performance.

Outcome	Adjusted difference	95% CI	F	*P*	Partial η^2
Immediate writing (T1)	5.03	[1.16, 8.89]	6.68	0.012	0.078
Delayed transfer (T2)	5.99	[2.31, 9.67]	10.49	0.002	0.117

Differences are defined as the GenAI-feedback class minus the teacher-feedback class; positive values indicate higher scores in the former. Models adjust for T0 writing scores; all F tests have df = (1, 79). The reported test of homogeneity of regression slopes was nonsignificant ($P = 0.412$). Unadjusted differences were 5.44 points at T1 (d = 0.35) and 6.40 points at T2 (d = 0.42).

The exploratory model showed an adjusted delayed-test advantage of 5.99 points for the AI-feedback class, 95% CI [2.31, 9.67], whereas the unadjusted independent-samples comparison was not statistically significant ($P = 0.061$). Because treatment condition and class membership were perfectly confounded, this model cannot identify the causal effect of feedback type, and its student-level *P* values cannot establish replicable class-level intervention effects (Table 4).

5.3. Associations of Feedback Literacy and SRL with Writing Transfer

To address RQ2, the associations of feedback literacy and SRL with delayed independent writing performance were examined.

Pretest-to-posttest change scores were computed for feedback literacy and SRL and correlated with T2 writing performance (Table 5). Change in feedback literacy was moderately correlated with T2 scores (r = 0.41, $P < 0.001$), whereas change in SRL showed a weaker positive correlation (r = 0.26, $P = 0.019$). Baseline feedback literacy and baseline SRL were more strongly correlated with T2 performance (r = 0.63 and 0.55, respectively). Changes in situated feedback-judgment accuracy and effective-revision proportion were also correlated with T2 scores (r = 0.33, $P = 0.003$, and r = 0.20, $P = 0.074$, respectively).

Table 5. Correlations of feedback literacy and SRL with delayed transfer scores.

Predictor	Pearson r with T2	95% CI	*P*
Baseline feedback literacy	0.634	[0.48, 0.75]	<0.001
Change in feedback literacy	0.413	[0.22, 0.58]	<0.001
Baseline SRL	0.548	[0.38, 0.68]	<0.001
Change in SRL	0.258	[0.04, 0.45]	0.019

Change = posttest minus pretest. Pearson's r is reported, with 95% CIs calculated using Fisher's z transformation; n = 82. Correlations for changes in situated judgment accuracy (r = 0.33) and effective-revision proportion (r = 0.20) are not included in the table.

A parsimonious regression model included baseline writing score, change in feedback literacy, and change in SRL as predictors of T2 performance (Table 6). Variance inflation factors ranged from 1.02 to 1.35, indicating no marked multicollinearity among the predictors.

Table 6. Regression analysis of delayed writing transfer scores.

Predictor	B	SE	β	95% CI	*P*
Baseline writing score	1.003	0.059	0.803	[0.89, 1.12]	<0.001
Change in feedback literacy	0.699	0.103	0.344	[0.49, 0.90]	<0.001
Change in SRL	0.197	0.107	0.093	[−0.02, 0.41]	0.070

The dependent variable was T2 writing score. B = unstandardized regression coefficient; β = standardized coefficient. $R^2 = 0.827$; adjusted $R^2 = 0.820$; $F(3, 78) = 124.0$, $P < 0.001$.

Feedback-literacy change was positively associated with delayed transfer performance after adjustment for baseline writing. The manuscript reports an additional 15.0% of explained variance for feedback-literacy change ($\Delta R^2 = 0.150$, $P < 0.001$). SRL change was weakly correlated with T2 performance in bivariate analysis (r = 0.26), but reportedly contributed only 0.7% additional explained variance after feedback-literacy change was included ($\Delta R^2 = 0.007$, $P = 0.070$). H2 therefore

received partial support: the feedback-literacy association was more robust, whereas the SRL association did not remain statistically significant in the multivariable model. No mediation pathway was tested.

A sensitivity regression examined whether these associations persisted after accounting for class and baseline learner characteristics (Table 7). After controlling for baseline writing, class membership, baseline feedback literacy, baseline self-regulated learning, and change in self-regulated learning, change in feedback literacy remained positively associated with delayed independent writing (B = 0.694, 95% CI [0.468, 0.920], $P < 0.001$). The class coefficient was not statistically significant (B = −0.109, $P = 0.949$), nor was change in self-regulated learning at the 0.05 threshold (B = 0.227, $P = 0.0503$). Adjusted R^2 was 0.817. The association between feedback literacy change and delayed writing remained after measured class and baseline differences were included; however, the two-class design precludes a causal interpretation.

Table 7. Multiple regression analysis of delayed writing transfer.

Predictor	Unstandardized B	95% CI	Standardized β	*P*
Baseline writing score (T0)	1.089	[0.867, 1.310]	0.872	<0.001
Class (AI = 1)	−0.109	[−3.522, 3.304]	−0.004	0.949
Baseline feedback literacy	−0.018	[−0.262, 0.227]	−0.012	0.886
Baseline self-regulated learning	−0.178	[−0.458, 0.102]	−0.086	0.209
Change in feedback literacy	0.694	[0.468, 0.920]	0.341	<0.001
Change in self-regulated learning	0.227	[−0.0003, 0.454]	0.107	0.0503

Note. Dependent variable: delayed transfer score (T2). B = unstandardized coefficient; β = standardized coefficient; 95% CIs are for B.

To contextualize the quantitative findings, all revision records were coded at the feedback-item level. A total of 738 feedback items were reported (three rounds per student, three items per round). Uptake and effective-revision rates were 68.3% and 34.4%, respectively, in the GenAI-feedback class, compared with 61.8% and 25.5% in the teacher-feedback class. The mean frequency of a "reasoned, appropriate rejection"—identifying an unsuitable suggestion and explaining why it should not be adopted—was 0.41 per student in the GenAI-feedback class and 0.32 in the teacher-feedback class. These descriptive process indicators suggest differences in both uptake and the quality of revision judgments.

6. Discussion

6.1. GenAI Feedback and Independent Writing Transfer

The AI-feedback class had a higher baseline-adjusted delayed-test mean than the teacher-feedback class, with an unadjusted standardized difference of d = 0.42, a small-to-moderate effect. A difference was already apparent at immediate posttest (d = 0.35) and was somewhat larger at delayed test, suggesting that the observed

pattern was not confined to the immediate writing assessment. One possible explanation is that reasons accompanying feedback helped students generalize individual corrections into rules they could later retrieve without tools. However, differences between classes may also reflect classroom climate, expectations, and other class-specific influences, and cannot be attributed to AI alone. Both classes showed some decline at delayed test, suggesting a need for periodic independent writing practice to reinforce learning.

The sensitivity regression further indicated that change in feedback literacy remained positively associated with delayed independent writing after controlling for class membership and initial writing, feedback literacy, and self-regulated learning. This suggests that the original association was not solely explained by average differences between the two classes or measured baseline characteristics. Nevertheless, because class and treatment were perfectly confounded and unmeasured factors could remain, the result supports only a relatively stable statistical association, not a causal or mediating role for feedback literacy.

6.2. The Interpretive Value of Feedback Literacy

Process data showed that situated-task judgment accuracy in the AI-feedback class increased from 52.0% to 77.0%, and the effective revision proportion from 45.5% to 76.6%, exceeding increases in the teacher-feedback class (52.8% to 60.2% and 42.1% to 53.5%, respectively). Both measures came from a situated judgment task administered at pretest (Week 1) and posttest (Week 8), each presenting 12 suggestions about the same essay. Students judged each suggestion and explained their decisions. The effective revision proportion used the 12 suggestions as its denominator, with the effectiveness criteria defined in Section 4.3.4. This measure differs from the item-level classroom effective revision rate in Section 5.3, which uses 369 classroom feedback items per class as its denominator. The former reflects judgment and revision under standardized conditions; the latter reflects overall revision output in ordinary instruction. For the teacher-feedback class, the classroom item-level effective revision rate was 25.5%, whereas situated-task effective revision proportions were 42.1% and 53.5% at pretest and posttest. Among 228 adopted classroom suggestions, 94 led to effective revisions, giving an effectiveness rate of 41.2% among adopted items. Because the tasks and denominators differ, the percentages are not directly comparable and do not establish equivalence of revision quality across measures.

Changes in situated judgment accuracy correlated more strongly with T2 ($r = 0.33$) than changes in self-regulated learning ($r = 0.26$), suggesting that evaluating and explaining feedback may be more closely related to transferable learning than mechanical uptake. Mean justified rejections were 0.41 and 0.32 per student in the two classes, a modest difference in the same direction. Some students could identify suggestions that sounded plausible but did not fit the text-behavior that uptake rates alone might incorrectly classify as insufficient engagement. Self-report and behavior did not always align: students reporting improved self-regulation

did not necessarily show comparable improvement in revision behavior. Social desirability and self-report bias therefore warrant consideration.

6.3. The Interpretive Value of Self-Regulated Learning

Learning reflections indicated that students in the GenAI-feedback class more frequently reported "making a revision plan before starting" after the second round and began referring to rules summarized in previous rounds by the third round (e.g., "check tense against time expressions" and "each paragraph needs a topic sentence"). Such reports are compatible with a sequence of planning, monitoring, and reflection, but they are self-reported, and keeping a learning log may itself cue these behaviors. They are therefore treated as contextual process evidence rather than proof that transfer occurred.

6.4. Limitations and Future Research

Each feedback condition was represented by only one intact class. Classroom culture, teacher expectations, and peer interaction were therefore confounded with the feedback condition, limiting causal inference and generalizability. The sample may have been underpowered to detect smaller associations; imperfect equivalence of writing tasks, variability in AI output, and self-report bias are additional concerns. Future research should replicate the study across multiple classes and schools, ideally with cluster-level randomization and repeated follow-up assessments.

7. Conclusion and Implications

This study evaluated independent writing transfer rather than the quality of AI-assisted revised drafts and used feedback judgments, revision behavior, and SRL reflections to contextualize classroom feedback processes. Following an eight-week intervention, the GenAI-feedback class had a 5.99-point higher baseline-adjusted delayed-test mean than the teacher-feedback class (95% CI [2.31, 9.67]; $d = 0.42$). This pattern co-occurred with changes in feedback-judgment accuracy and effective revision. However, the one-class-per-condition design limits the findings to associations observed in these specific classrooms. The sample may also have been insufficient to detect effects smaller than $d = 0.30$ reliably, and equivalence of the three writing prompts was only preliminarily assessed. In lower secondary English instruction, feedback activities may usefully encourage students to explain suggestions, compare alternatives, record revision rationales, and periodically complete unsupported writing tasks, rather than treating external feedback as evidence of acquired competence.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Carless, D., & Boud, D. (2018). The Development of Student Feedback Literacy: Enabling

Uptake of Feedback. *Assessment & Evaluation in Higher Education, 43,* 1315-1325. https://doi.org/10.1080/02602938.2018.1463354

Ekizoğlu, M., & Demir, A. N. (2025). The Role of AI Assisted Writing Feedback in Developing Secondary Students Writing Skills. *Discover Education, 4,* Article No. 454. https://doi.org/10.1007/s44217-025-00919-3

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F. et al. (2023). ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences, 103,* Article ID: 102274. https://doi.org/10.1016/j.lindif.2023.102274

Parra Núñez, R., & Castrillo de Larreta-Azelain, M. D. (2025). Impacto de ChatGPT como herramienta de retroalimentación en la escritura en inglés como lengua extranjera: Revisión sistemática. *Revista Internacional de Lenguas Extranjeras, No. 23,* 95-119.

Perkins, D. N., & Salomon, G. (1992). Transfer of Learning. In T. Husén, & T. N. Postlethwaite (Eds.), *The International Encyclopedia of Education* (2nd ed., pp. 6452-6457). Pergamon Press.

Sulistiyo, U., Angraini, D., & Fikri, H. M. (2026). GenAI-Mediated Feedback for EFL Writing: Effects, Student Uptake and Ethics: A Systematic Review. *Learning Futures and Emerging Technologies,* 1-16. https://doi.org/10.1108/lfet-04-2026-0052

Zimmerman, B. J. (2002). Becoming a Self-Regulated Learner: An Overview. *Theory Into Practice, 41,* 64-70. https://doi.org/10.1207/s15430421tip4102_2

Appendix A. English Writing Test Prompts and Scoring Rubric

Three narrative tasks were administered at pretest (T0), immediate posttest (T1), and delayed test (T2). Each required an 80 - 100-word submission to an English-language school magazine, with topics covering school life, weekend activities, and personal growth to reduce practice effects. All tests were timed and completed independently without AI prompts, dictionaries, or peer discussion.

(A) Writing Tasks

T0 (Pretest): Write "A Memorable School Day" for the English school magazine. Describe what happened, how you felt, and what you learned. Suggested length: 80 - 100 words.

T1 (Immediate posttest): Write "A Meaningful Weekend Activity" for the English school magazine. Explain the activity, what you did, how you felt, and why it mattered. Suggested length: 80 - 100 words.

T2 (Delayed test): Write "An Experience That Helped Me Grow." Describe an experience, a problem you encountered, how you addressed it, and what you learned. Suggested length: 80 - 100 words.

(B) Analytic Scoring Rubric

Dimension (weight)	Level 4	Level 3	Level 2	Level 1
Task fulfillment (25%)	All three requirements fully addressed with sufficient detail	Most requirements addressed; limited detail	Requirements only partly met; some off-topic content	Substantially off topic or too little content
Organization and coherence (25%)	Clear structure and natural transitions	Complete structure; generally smooth transitions	Loose structure and awkward transitions	No clear structure; disconnected sentences
Vocabulary (20%)	Accurate, varied vocabulary with some effective choices	Mostly accurate but repetitive vocabulary	Limited vocabulary and collocation errors	Frequent lexical errors impede understanding
Grammar (20%)	Varied structures with very few errors	Mostly simple sentences with a few errors	Numerous errors without seriously impairing meaning	Pervasive errors impede understanding
Appropriateness (10%)	Clear awareness of register and audience	Generally appropriate register	Inconsistent register	Inappropriate for the required genre

Note. Each dimension is rated on a four-level scale. Total score = Σ (level score ÷ 4 × maximum points for the dimension), converted to a 100-point scale for analysis. Raters must complete anchor-script training and meet the reliability criterion (ICC ≥ 0.75) before formal scoring.

(C) Scoring Procedures

Two English teachers completed calibration training using anchor scripts (two sample essays per performance level). They independently scored essays without knowledge of class or test occasion. The mean of the two weighted total scores was used. Before formal scoring, 20 essays were double-scored to check consistency; discrepancies were discussed using the rubric and anchor examples.

Appendix B. Feedback Literacy Questionnaire

Instructions

Dear student, the following 20 statements describe what you think and do when receiving feedback on English writing. Choose the response that best reflects your usual experience: ① Strongly disagree ② Disagree ③ Unsure ④ Agree ⑤ Strongly agree. This questionnaire is unrelated to your English grades; there are no right or wrong answers. Please answer independently and select the option that best reflects your own views.

Item	Dimension	Statement
1	Feedback comprehension	I believe comments from teachers or AI help me notice problems in my writing that I had overlooked.
2		I can identify which part of my writing a comment refers to (such as the introduction, a sentence, or a word).
3		I know feedback does not only identify errors; it can also point out what I did well.
4		When I receive feedback, I can generally tell whether it concerns content or language.
5		I believe revising my writing more than once is necessary for improving my writing skills.
6		I can distinguish errors that must be corrected from aspects that could be improved.
7		I compare feedback with the scoring criteria for content, organization, and language.
8	Feedback judgment	Before adopting an AI suggestion, I consider whether it is correct.
9		I can recognize suggestions that distort my intended meaning or make a sentence read awkwardly.
10		If I do not understand feedback, I ask my teacher or classmates rather than ignoring it.
11		I can judge whether a suggestion suits the purpose of my essay.
12		I compare feedback from AI, teachers, and classmates to see whether they agree.
13		I know AI sometimes gives suggestions that sound reasonable but do not fit my essay.
14		Whenever AI or a teacher suggests a change, I make it without much thought about its suitability.
15	Feedback use	I plan how to revise before making changes rather than revising impulsively.
16		When revising, I reread the whole essay to check whether changes affect coherence.

Continued

17	I note the problems identified in feedback to remind myself next time.
18	After revising, I check whether I have addressed each feedback point.
19	I apply what I learn from feedback to my next essay.
20	I voluntarily explain to my teacher or classmates why I made particular revisions.

Note. Items 1 - 7 assess feedback comprehension, Items 8 - 14 feedback judgment, and Items 15 - 20 feedback use. Item 14 is reverse-scored; the other 19 items are positively scored.

Scoring and Evaluation Rules:

(1) Response scoring. Options ① - ⑤ receive scores of 1 - 5. Item 14 is reverse-scored as 6 - raw score before calculating dimension scores; the remaining 19 items are scored positively.

(2) Dimension scores. Feedback comprehension = sum of Items 1 - 7 (7 - 35); feedback judgment = sum of Items 8 - 14 (7 - 35); feedback use = sum of Items 15 - 20 (6 - 30).

(3) Total score. Sum of all 20 items (20 - 100). Higher scores indicate greater feedback literacy. Total and dimension scores were analyzed as continuous variables; no pass threshold or arbitrary high/medium/low categories were imposed.

(4) Missing data. If no more than one item was missing within a dimension, the rounded mean of completed items in that dimension was substituted. A questionnaire with two or more missing items in a dimension was excluded. No missing values occurred at either assessment.

(5) Reliability. In the study sample of 82 students, Cronbach's α was .88 for the total scale and 0.76 - 0.81 for the three dimensions (see table). No items required removal.

Dimension	No. of items	Item	Cronbach's α
Feedback comprehension	7	Items 1 - 7	0.81
Feedback judgment	7	Items 8 - 14	0.79
Feedback use	6	Items 15 - 20	0.76
Total scale	20	Items 1 - 20	0.88

Note. Cronbach's α values were calculated using baseline data from the study sample (n = 82). Dimension scores are sums rather than means.

Validity Evidence:

(1) Item development. Items were developed from Carless and Boud's (2018) feedback literacy framework, emphasizing appreciation of feedback, judgment, and action, and organized into feedback comprehension, judgment, and use.

Three to five behavioral indicators were specified for each dimension and converted into single-statement items. The initial pool contained 26 items.

(2) Expert review. Five experts (two English teaching researchers, two senior secondary-school teachers, and one associate professor of educational measurement) conducted two rounds of review covering dimension fit, clarity, relevance, and response burden. First-round item-level content validity indices (I-CVI) ranged from 0.80 to 1.00 and S-CVI/Ave was 0.93. Three items were deleted, three combined, and six reworded; the second round confirmed I-CVI ≥ 0.80 for all 20 pilot items.

(3) Pilot and item analysis. The questionnaire was piloted in three Grade 8 classes at another school, with 110 distributed and 104 valid responses. Item critical ratios ranged from 3.05 to 9.42 ($P < 0.01$), and item-total correlations from 0.38 to 0.72, meeting retention criteria.

(4) Structural validity. Exploratory factor analysis of the pilot sample (n = 104) yielded KMO = 0.86 and Bartlett's test $\chi^2(190) = 874.3$ ($P < 0.001$). Three factors explained 58.7% of total variance, with loadings of 0.52 - 0.79 and cross-loadings below 0.40, consistent with the proposed dimensions. Confirmatory factor analysis was not performed in the formal sample (n = 82); instead, interdimension correlations (0.42 - 0.55) and dimension-total correlations (0.78 - 0.86) were reported as supplementary internal-structure evidence.

(5) Scope. Test-retest reliability and criterion-related validity were not examined; findings should be interpreted alongside situated tasks and revision records.

Appendix C. Self-Regulated Learning Questionnaire

Instructions

Dear student, the following 15 statements describe your approaches and habits in English writing. Based on your experience during the past month, select the option that best describes you: ① Strongly disagree ② Disagree ③ Unsure ④ Agree ⑤ Strongly agree. This questionnaire does not affect your English grade, and there are no right or wrong answers. Please complete it independently.

Item	Dimension	Statement
P1	Goal planning	Before writing, I decide what main ideas I want to express.
P2		Before writing, I set goals for what I need to improve.
P3		I allow time to check my essay instead of submitting it immediately.
P4		Before writing, I review the task requirements and scoring criteria.
P5		I anticipate where I am most likely to make mistakes.
M1	Process monitoring	While writing, I check whether my sentences express what I mean.
M2		When I receive a revision suggestion, I first consider whether it suits my essay.

Continued

M3		When I do not understand a suggestion, I try to find an explanation.
M4		While writing, I pause to reread what I have written.
M5		I often finish my essay and submit it without rereading or checking.
R1	Self-reflection	After revising, I compare my first and revised drafts.
R2		I summarize the problems that occurred most often in my writing.
R3		In my next essay, I try to use strategies I summarized previously.
R4		I reflect on which revisions actually solved problems and which did not.
R5		Before submitting, I consider whether this essay is better than my previous one.

Note. M5 is reverse-scored; the other 14 items are positively scored. P, M, and R denote goal planning, process monitoring, and self-reflection, respectively.

Scoring and Evaluation Rules:

(1) Response scoring. Options ① - ⑤ receive scores of 1 - 5. M5 is reverse-scored as 6 - raw score; the other 14 items are positively scored.

(2) Dimension scores. Goal planning = sum of P1 - P5 (5 - 25); process monitoring = sum of M1 - M5 (5 - 25); self-reflection = sum of R1 - R5 (5 - 25).

(3) Total score. Sum of all 15 items (15 - 75). Higher scores indicate stronger self-reported writing self-regulation. Total and dimension scores were treated as continuous, without pass thresholds or arbitrary categories.

(4) Missing data. When at most one item was missing within a dimension, the rounded mean of completed items was substituted. A questionnaire with two or more missing items in a dimension was excluded. No missing values occurred at either assessment.

(5) Reliability. In the study sample (n = 82), Cronbach's α was 0.85 for the total scale and 0.75 - 0.78 for its three dimensions (see table). No items required removal.

Dimension	No. of items	Item	Cronbach's α
Goal planning	5	P1 - P5	0.78
Process monitoring	5	M1 - M5	0.76
Self-reflection	5	R1 - R5	0.75
Total scale	15	P1 - R5	0.85

Validity Evidence:

(1) Item development. Items were developed from Zimmerman's (2002) cyclical model of self-regulated learning, encompassing forethought, performance, and self-reflection, and mapped onto goal planning, process monitoring, and self-re-

flection. Behavioral indicators were drafted for each phase and converted into single-statement items. The initial pool contained 18 items.

(2) Expert review. The same five experts and two review rounds used for the feedback literacy scale assessed the same criteria. I-CVI values ranged from 0.80 to 1.00 and S-CVI/Ave was 0.91. Two items were deleted, one combined, and five reworded; the second round confirmed I-CVI ≥ 0.80 for all 15 pilot items.

(3) Pilot and item analysis. The questionnaire was administered in the same pilot (104 valid responses). Critical ratios ranged from 3.12 to 8.87 ($P < 0.01$), and item-total correlations from 0.35 to 0.69, meeting retention criteria. The correlation of reverse-worded Item M5 with other items in its dimension was −0.31 before reverse scoring and 0.44 afterward, as reported in the source manuscript.

(4) Structural validity. Exploratory factor analysis of pilot data (n = 104) yielded KMO = 0.83 and Bartlett's test $\chi^2(105) = 612.5$ ($P < 0.001$). Three factors explained 61.2% of total variance, with loadings of 0.55 - 0.81 and cross-loadings below .40, consistent with the proposed dimensions. In the formal sample (n = 82), interdimension correlations ranged from 0.39 to 0.52 and dimension-total correlations from 0.75 to 0.84 (all $P < 0.001$).

(5) Scope. As with the feedback literacy questionnaire, test-retest reliability and criterion-related validity were not examined. Writing self-regulation is context-dependent, and the questionnaire captures students' general perceptions of their writing habits; results should be compared with reflection records and revision behavior.