

Research on English Writing Teaching Reform Based on Multi-subject Collaborative Evaluation: A Case Study of the Intelligent English Writing Course

Zhiying Li

South China Business College, Guangdong University of Foreign Studies

Author Note

The author declares no conflicts of interest to disclose. Correspondence concerning this article should be addressed to Zhiying Li, School of English Language and Culture, South China Business College, Guangdong University of Foreign Studies, Guangzhou, Guangdong 510545, China. Email: 213141@gwng.edu.cn

Abstract

Driven by the new-era educational evaluation reform emphasizing process-based and developmental assessment, traditional college English writing teaching is constrained by teacher-dominated single evaluation, delayed feedback, and low student participation, which hinder the sustainable improvement of students' writing competence. While artificial intelligence (AI) has facilitated innovative writing evaluation approaches, most existing studies merely adopt single AI or conventional teacher-peer evaluation, with few empirical explorations on progressive multi-subject evaluation models suitable for application-oriented universities. To fill this gap, this study takes the Intelligent English Writing course as the research object and constructs a four-stage progressive collaborative evaluation model integrating student self-evaluation, peer evaluation, critical AI optimization, and whole-process teacher diagnosis. A one-semester teaching intervention was conducted on four parallel classes. Pre-test and post-test writing scores as well as valid questionnaire data ($N=59$, Cronbach's $\alpha=0.89$) were used to examine its effectiveness. The results demonstrate that the proposed model significantly improves students' writing performance, learning engagement, metacognitive reflection, and critical feedback screening abilities, while effectively reducing their writing anxiety and optimizing the traditional classroom ecology. This study clarifies the auxiliary role and critical application principles of AI tools, balances technological empowerment and humanistic teaching, and establishes a replicable practical paradigm for intelligent writing instruction. It provides credible evidence and actionable references for college English writing teaching reform and optimized educational evaluation.

Received 18 April 2026; revised 28 May 2026; accepted 16 June 2026; available online 25 June 2026; version of record 30 June 2026.

Citation: Li, Z. Y. (2026). Research on English writing teaching reform based on multi-subject collaborative evaluation: A case study of the Intelligent English Writing course. *Journal of Language*, 2(2), 53–73. <https://doi.org/10.64699/26LVQM6056>

Keywords: Intelligent English Writing course, multi-subject collaborative evaluation, formative assessment, AI-assisted teaching, teaching reform

1. Introduction

In recent years, China has comprehensively promoted the reform of the educational evaluation system, aiming to break the single score-oriented evaluation mechanism and construct a scientific and diversified evaluation system. The *Overall Plan for Deepening Educational Evaluation Reform in the New Era* explicitly proposes to reverse the “score-only” tendency and prioritize process-based and developmental evaluation (Central Committee of the Communist Party of China & State Council, 2020). Guided by this policy, university teaching has undergone a fundamental shift from outcome-oriented evaluation to a comprehensive evaluation system centered on whole-learning process and competency development. This reform is not merely an adjustment of evaluation methods, but a systematic transformation of educational concepts and talent cultivation models, which fundamentally promotes in-depth changes in teaching and learning behaviors through the reconstruction of evaluation mechanisms.

This transformative shift is particularly urgent in foreign language education. As an application-oriented undergraduate institution, South China Business College, Guangdong University of Foreign Studies focuses on cultivating students’ practical language application competence. Nevertheless, traditional English writing teaching still relies heavily on teachers’ summative scoring, characterized by low student participation and delayed feedback. Such a rigid evaluation model fails to support the continuous improvement of students’ writing ability, restricts students’ learning initiative, and hinders the further development of foreign language teaching reform.

With the rapid development of artificial intelligence (AI) technology, innovative possibilities have emerged for educational evaluation. Writing feedback tools based on natural language processing (NLP) technology can conduct multi-dimensional textual analysis and provide real-time instructional support for the learning process (Shaik et al., 2022). In the field of intelligent technology-enabled language education, generative AI is profoundly reshaping the paradigm of language teaching and evaluation. Reasonably defining the role of AI tools and constructing human-machine collaborative teaching models has become a core research hotspot in contemporary language education studies (Yue et al., 2026).

However, most existing studies only focus on single AI evaluation or traditional teacher-student mutual evaluation, with insufficient exploration of the progressive multi-subject collaborative evaluation system integrating student self-evaluation, peer evaluation, AI intelligent evaluation, and teacher final evaluation. Empirically verified and promotable curriculum practice paradigms tailored to application-oriented undergraduate colleges are still scarce. Against this research gap, this study constructs a four-stage progressive multi-subject collaborative evaluation model based on the teaching practice of the Intelligent English Writing course. Through empirical teaching experiments, this research verifies the effectiveness of the reform, solves the inherent drawbacks of traditional writing evaluation, and provides practical references for optimizing college foreign language writing teaching and implementing educational evaluation reform.

2. Theoretical Foundations

This study integrates two mainstream foreign language teaching theories, namely

formative assessment theory and multi-subject collaborative evaluation theory. Combined with intelligent teaching scenarios, it constructs a targeted collaborative evaluation system for the course.

2.1 Formative Assessment Theory

Formative assessment embeds evaluative behaviors into the whole learning process, replacing single summative outcome assessment with continuous, diagnostic, and developmental feedback (Black & Wiliam, 1998). It guides students to optimize learning behaviors, improve knowledge systems, and enhance comprehensive competencies, thereby realizing the core goal of “promoting learning through assessment”. Recent scholarship further reinforces the intrinsic value of formative assessment in higher education teaching and learning reform. Rutherford et al. (2025) emphasize that assessment and corresponding feedback serve as powerful drivers of student learning development, highlighting a prevalent sector shift from traditional summative assessment of learning to advanced formative assessment for learning and even assessment as learning. This paradigm shift underscores that assessment is an integral component of the learning process rather than a separated concluding procedure, playing a decisive role in fostering students’ self-regulated learning abilities and learner independence. To maximize instructional effectiveness, formative assessment practices must be student-centered, contextually integrated, authentic, and engaging, enabling personalized competence development and empowering sustainable learner growth (Rutherford et al., 2025). The classic research of Black and Wiliam (1998) proved that high-quality and regular classroom process feedback could effectively optimize learners’ learning strategies and improve learning outcomes, serving as a core driving force

for classroom teaching improvement.

In the context of intelligent foreign language teaching, AI writing feedback powered by natural language processing (NLP) technology further complements and empowers formative assessment implementation. It can realize standardized and real-time evaluation of text grammar, vocabulary, sentence patterns, and discourse structure, presenting prominent advantages in grammatical error correction, formatting standardization, and basic textual optimization (Stevenson & Phakiti, 2014). Such intelligent feedback effectively compensates for the limitations of traditional teacher evaluation, including heavy teaching workload and delayed manual feedback, enabling more timely and frequent process-based guidance that aligns perfectly with the core connotation of formative assessment. Nevertheless, international academic consensus indicates that AI feedback has inherent deficiencies: it lacks in-depth contextual interpretation and humanistic connotation and cannot accurately identify learners' internal logical thinking and emotional expression in writing. Therefore, AI tools can only function as auxiliary educational resources, and their practical instructional effectiveness heavily depends on learners' critical screening, rational reflection, and contextual adaptive application (Qin, 2026). This theoretical cognition underpins the core principle of "critical use of AI" adopted in this study and standardizes the whole-process intelligent feedback application procedure in multi-round writing iteration.

Different from traditional summative assessment that only judges final learning outcomes, formative assessment focuses more on empowering the whole learning process and attaching importance to students' growth trajectory and competency progression. It highly conforms to the developmental orientation of new-era educational evaluation reform and

provides solid theoretical support for the construction of the whole-process progressive evaluation model in this study.

2.2 Multi-subject Collaborative Evaluation Theory

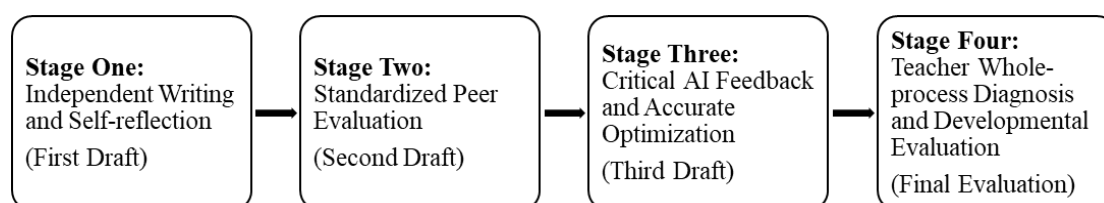
The theoretical origin of multi-subject collaborative evaluation can be traced to the fourth-generation evaluation theory proposed by Guba and Lincoln (1989). Differing from traditional single-dimensional and objective evaluation paradigms that merely focus on measurable results, this constructivist evaluation theory advocates that educational evaluation is a dynamic, negotiated, and co-constructed process involving multiple stakeholders. It emphasizes responsiveness to diverse evaluation demands, integration of multi-party perspectives, and abandonment of the single authoritative evaluation standard dominated by educators, laying the fundamental theoretical foundation for multi-subject participatory evaluation in education. On this basis, multi-subject evaluation theory further breaks the limitation of the single teacher-centered evaluation model. It advocates integrating diverse evaluators, evaluation criteria, and evaluation methods to comprehensively, objectively, and stereoscopically reflect learners' learning status and comprehensive ability level (Liu & Carless, 2006). In foreign language writing teaching, student self-evaluation activates metacognitive awareness and helps learners independently identify personal writing deficiencies. Peer evaluation complements learning from peer perspectives, deepens students' understanding of writing criteria, and cultivates their critical thinking and textual analysis competence.

Existing international studies have confirmed that multi-subject interactive evaluation can effectively make up for the one-sidedness of single evaluation. However, most previous

studies failed to realize the staged and progressive integration of multiple evaluators, resulting in fragmented evaluation processes and low feedback utilization efficiency. These deficiencies leave sufficient room for the model innovation of this research.

3. Teaching Practice Paths

Based on the Intelligent English Writing course of an application-oriented undergraduate college, this study selects four parallel freshman classes as research objects. Two classes with 26 and 35 students respectively are assigned as the experimental group, while the other two classes of 27 and 26 students form the control group. A four-stage progressive multi-subject collaborative evaluation model is constructed, including student self-evaluation for the first draft, peer evaluation for the second draft, critical AI optimization for the third draft, and teacher whole-process diagnostic final evaluation. Standardized evaluation criteria, implementation procedures, revision requirements, and reflection norms are formulated to form a complete teaching closed loop of “writing-evaluation-reflection-iteration”, fully implementing the concept of process-based and developmental evaluation. The specific practice paths are as follows:



3.1 Stage One: Independent Writing and Self-reflection (First Draft)

The core objective of this stage is to restore students' real writing level and activate their independent reflection awareness. Teachers assign theme-based writing tasks and require students to complete the first draft independently without any external auxiliary tools,

so as to fully present their real language reserve, logical thinking, and writing weaknesses and avoid distorted writing levels caused by external intervention. After completing the first draft, students are required to finish standardized self-reflection reports in accordance with the authoritative and empirically validated diagnostic writing evaluation framework proposed by Knoch (2009). This study adopts a hierarchical writing assessment system with three primary dimensions and nine refined secondary indicators to evaluate EFL writing performance, covering linguistic competence, textual organization, and task fulfillment. First, the linguistic competence dimension consists of three sub-indicators: accuracy (the severity and frequency of grammatical and spelling errors in writing), repair fluency (the proportion of self-correction, restatement and erasure in texts, which reflects the fluency of on-site writing output), and lexical complexity (the richness of word choice and the appropriateness of academic vocabulary application). Second, the textual organization dimension includes paragraphing (the rationality of paragraph division and the clarity of thematic focus in each paragraph), coherence (the overall logical fluency and readability of the full text), and cohesion (the richness and appropriateness of cohesive devices such as conjunctions, pronouns, and structural substitutions). Third, the task fulfillment dimension involves data description (the accurate presentation of core chart data, changing trends and comparative relationships), data interpretation (reasonable analysis and logical inference of underlying rules and causes behind the statistical data), and topic-based extended discussion, including targeted derived conclusions, practical implications and innovative expansions corresponding to specific writing requirements. This standardized and validated multi-layered assessment system enables students to objectively identify personal writing strengths, existing problems,

and preliminary revision ideas during self-evaluation.

Teachers do not participate in pre-stage intervention and evaluation, but only collect students' first drafts and reflection texts to retain original learning data, providing a baseline for subsequent multi-round iteration comparison and effect analysis.

3.2 Stage Two: Standardized Peer Evaluation (Second Draft)

This stage aims to deepen students' understanding of writing evaluation criteria and improve their textual analysis and critical thinking competence through peer evaluation. Teachers conduct special training on peer evaluation in advance to unify scoring standards, feedback expressions, and revision norms based on Knoch's (2009) validated multi-layer diagnostic writing assessment framework. The standardized evaluation system covers three primary dimensions and nine secondary indicators: linguistic competence (accuracy, repair fluency, lexical complexity), textual organization (paragraphing, coherence, cohesion), and task fulfillment (data description, data interpretation, and topic-based extended discussion). Each sub-indicator is clearly defined to ensure objective, targeted, and consistent peer judgment. The unified and detailed evaluation dimensions help eliminate evaluation quality imbalance caused by individual student ability differences and ensure consistent, objective, and targeted peer feedback. After training, an anonymous cross-peer evaluation mode is adopted. Students exchange first drafts, score items by item according to unified standards, mark existing problems, and put forward targeted revision suggestions, focusing on evaluating humanistic and logical problems that cannot be accurately identified by AI tools.

After peer evaluation, students sort out peer feedback, critically screen effective suggestions and abandon unreasonable ones based on their own writing intentions and

evaluation standards, complete the revision of the second draft, and write peer reflection records to illustrate the basis for adopting or discarding feedback suggestions.

3.3 Stage Three: Critical AI Feedback and Accurate Optimization (Third Draft)

As the core innovative link of this model, this stage solves the prominent problem of blind adoption and over-reliance on traditional AI evaluation by implementing the critical application of AI tools. Students input the revised second draft into the intelligent writing evaluation tool to obtain standardized feedback strictly based on Knoch's (2009) empirically validated multi-dimensional diagnostic writing assessment framework.

Different from the traditional mode of directly applying all AI feedback, this study requires students to screen AI suggestions item by item in combination with writing context and thematic conception, strictly referring to Knoch's (2009) hierarchical evaluation dimensions covering linguistic competence, textual organization, and task fulfillment. Students retain and optimize AI revisions that effectively improve writing accuracy, fluency, lexical richness, textual logic and task completion quality, while discarding rigid, context-inconsistent suggestions that damage thematic coherence and personalized expression. Students are required to record the whole process of screening, revision, and abandonment in detail, form the optimized third draft, and write Critical Reflection Logs on AI Feedback with specific revision cases, so as to consolidate critical thinking and eliminate blind dependence on intelligent tools.

3.4 Stage Four: Teacher Whole-process Diagnosis and Developmental Evaluation (Final Evaluation)

This stage focuses on conducting comprehensive and developmental evaluation based

on students' whole-process learning performance. Abandoning the traditional single final-score evaluation mode, teachers integrate students' multi-version writing texts (first, second, and third drafts), stage reflection reports, peer evaluation records, and AI feedback application logs to carry out whole-process tracking evaluation. The evaluation focus is no longer limited to final text quality, but covers four core dimensions: the progressive improvement of writing content and linguistic competence, the ability to screen and transform peer and AI feedback, the depth of independent reflection and problem analysis, and the developmental changes in writing thinking and learning attitudes.

Based on whole-process diagnostic results, teachers provide personalized growth comments and targeted improvement plans for each student, realizing one-on-one academic guidance and fully achieving the teaching goal of promoting learning and development through evaluation.

4. Practical Effect Analysis

This study selects four parallel classes with no significant difference in academic foundation in the same grade for a one-semester teaching practice. The experimental group consists of two classes with 26 and 35 students respectively and adopts the multi-subject collaborative evaluation model. The remaining two classes, with 27 and 26 students in turn, serve as the control group using the traditional teacher-centered summative evaluation model. The teaching effects are analyzed from three dimensions: objective academic performance, subjective competency improvement, and learning behavioral changes based on quantitative data and questionnaire surveys.

4.1 Improvements in Students' Writing Performance

Independent sample t-tests were conducted to compare the pre-test and post-test writing scores of the two groups, so as to objectively verify the effectiveness of the multi-subject collaborative evaluation teaching model. The statistical analysis was based on valid data from 61 students in the experimental group and 53 students in the control group, ensuring the authenticity and representativeness of the experimental results. All writing scores involved in this study were automatically scored by the Pigai writing evaluation platform. The results show that before the teaching practice, there is no significant difference in the average writing scores between the experimental group and the control group ($P>0.05$), indicating consistent basic learning levels and valid experimental grouping.

After the one-semester teaching intervention, both groups achieved certain progress in English writing performance, but the improvement degree varied greatly. Specifically, the experimental group adopting the multi-subject collaborative evaluation model obtained far more significant score growth than the control group with traditional summative evaluation. The average writing score of the experimental group increased by 8.72 points, with the excellent rate rising by 21.30% and the pass rate increasing by 6.56%. In contrast, the control group only achieved a slight average score improvement of 2.23 points, accompanied by mild growth in excellent rate (7.55%) and passing rate (3.77%).

Table 1 Summary of Pre-test and Post-test English Writing Scores of the Experimental Group and Control Group

Group	Pre-test Mean Score	Post-test Mean Score	Mean Score Improvement
Experimental Group	70.35	79.07	8.72
Control Group	69.89	72.12	2.23

Table 2 Improvement in Excellence Rate and Passing Rate of Writing Scores in the Experimental and Control Group

Group	Indicator	Pre-test Value	Post-test Value	Improvement
Experimental Group	Excellence Rate (≥ 85 points)	11.48%	32.78%	21.30%
	Passing Rate (≥ 60 points)	81.96%	88.52%	6.56%
Control Group	Excellence Rate (≥ 85 points)	11.32%	18.87%	7.55%
	Passing Rate (≥ 60 points)	83.02%	86.79%	3.77%

Further independent-samples t-test results confirmed the significant promotional effect of the proposed teaching model. The pre-test t-value was 0.24 ($P > 0.05$), which further verified the homogeneity of the two groups before the experiment. After the intervention, the post-test t-value reached 3.28 ($P < 0.01$), demonstrating that the multi-subject collaborative evaluation model produced statistically significant positive effects on students' writing ability improvement.

Table 3 Independent-samples T-test of Pre-test and Post-test Writing Scores Between Two Groups

Test Stage	Group	t	p
Pre-test	Experimental Group	0.24	> 0.05
	Control Group		
Post-test	Experimental Group	3.28	< 0.01
	Control Group		

The most significant improvements of the experimental group are fully reflected in the three validated evaluation dimensions proposed by Knoch (2009). In terms of linguistic competence, students achieved prominent progress in grammatical accuracy, writing repair fluency, and lexical complexity, with fewer grammatical and spelling errors, smoother on-site writing output, and more appropriate use of diversified academic vocabulary. In terms of textual organization, students showed enhanced rationality of paragraph division, higher global textual coherence, and more flexible and accurate application of various cohesive devices. Regarding task fulfillment, students performed better in accurate data description, in-depth data interpretation, and reasonable topic-based extended discussions, realizing comprehensive and in-depth task completion. In contrast, the control group showed no obvious fluctuation in overall writing performance, with only marginal improvement in basic linguistic accuracy and no substantial progress in textual organization and task fulfillment. The quantitative data verify that the multi-subject collaborative evaluation model can effectively promote the systematic progression of students' writing ability and has prominent advantages over the traditional single evaluation mode.

4.2 Enhancement of Learning Engagement and Metacognitive Competence

This study adopts the self-designed Questionnaire on English Writing Learning Status to conduct an anonymous survey on 61 students in the experimental group. A total of 59 valid questionnaires are recovered, with an effective recovery rate of 96.72%. The questionnaire covers four core dimensions (learning engagement, independent reflection competence, feedback acceptance and screening competence, and learning motivation) and consists of 20 five-point Likert scale items. The reliability test shows that the overall Cronbach's α

coefficient is 0.89, indicating excellent questionnaire reliability and valid statistical analysis value. The detailed statistical results of each dimension are shown in Table 4.

Table 4 Statistical Results of Students' Learning Status Questionnaire

Dimension	Mean Score	Positive Recognition Rate
Learning Engagement	3.78	70.17%
Independent Reflection	3.86	76.27%
Competence		
Feedback Screening &	3.81	73.56%
Acceptance Competence		
Learning Motivation &	3.69	68.98%
Comprehensive Improvement		

Note. 1. A total of 20 valid scale items are included in the survey, with 5 items for each dimension (N=59); 2. Positive recognition rate refers to the total proportion of respondents choosing “agree” and “strongly agree”; 3. Dimension mean scores are weighted averages of measured scores of corresponding items, all derived from real survey data.

The data of 20 scale items show that the mean score of all dimensions is higher than 3.5, indicating a generally high recognition of the multi-subject collaborative evaluation model among students. In terms of learning engagement, 70.17% of students can actively participate in the whole process of writing, revision and reflection, getting rid of the passive learning state in traditional writing courses. The multi-round iteration mode effectively activates students' classroom participation and autonomous learning enthusiasm.

For independent reflection competence, 76.27% of students can accurately grasp

English writing evaluation criteria through self-evaluation, peer evaluation and multi-version text comparison, and actively analyze personal deficiencies in discourse structure and logical expression, gradually forming the autonomous learning habit of self-inspection and self-optimization after writing. In terms of feedback screening ability, more than 70% of students develop critical learning awareness. They no longer blindly follow teacher and AI feedback, but rationally screen and select revision suggestions based on text context and writing themes, effectively avoiding AI dependence and significantly improving textual discrimination and critical thinking ability. Nevertheless, nearly 30% of the students admit that they still rely heavily on AI tools when revising their writing.

In terms of learning motivation, the whole-process visualized iteration evaluation effectively alleviates students' writing anxiety and enhances their learning confidence. Students' comprehensive abilities in language expression and logical construction are substantially improved, and their intrinsic motivation to optimize writing competence is continuously strengthened. Overall, the measured data fully prove that the multi-subject collaborative evaluation model optimizes students' writing learning status and significantly improves their metacognitive ability and autonomous learning literacy.

4.3 Changes in Classroom Teaching Ecology

The multi-subject collaborative evaluation model completely subverts the traditional rigid classroom ecology of "one-way teacher correction and passive student acceptance". Students transform from task-oriented completers to whole-process participants, reflectors and optimizers of writing learning, which significantly enhances classroom interaction and inquiry. Teachers shift from single scorers and correctors to instructional guides, learning

condition diagnosticians and competency cultivators, greatly improving the pertinence and accuracy of teaching.

Meanwhile, multi-version writing texts and whole-process reflection records realize the visualization of learning processes, providing solid data support for the implementation of process-based evaluation and personalized teaching. The teaching transformation from “result-oriented evaluation” to “process-empowered evaluation” is successfully realized.

5. Problems and Reflections

Combined with one-semester teaching data and classroom observation, this study summarizes the practical problems in the operation of the model, analyzes the underlying causes, and proposes targeted optimization strategies.

5.1 Excessive AI Dependence and Stratified Critical Application Ability

The teaching practice finds that students with weak academic foundations are prone to blind dependence on AI feedback. To save revision time, some students directly apply all AI-generated sentence patterns, vocabulary and paragraph revisions without adapting to personal writing context, resulting in disjointed texts with elegant sentences but deviated themes. For instance, some students mechanically adopt AI’s complex sentence replacement suggestions in argumentative writing, leading to stiff expressions and logical gaps. This phenomenon is mainly attributed to low-level learners’ insufficient language discrimination and contextual control ability, as well as ambiguous cognition of AI’s auxiliary positioning, which hinders their accurate and critical application of AI tools.

5.2 Uneven Peer Evaluation Quality and Insufficient Standardization

Affected by individual differences in language proficiency, critical thinking and

learning attitudes, peer evaluation quality varies greatly. High-level students can provide accurate and targeted feedback to assist peer optimization, while underachieving students can only identify superficial grammatical errors but fail to judge in-depth problems in discourse logic and thematic expression, resulting in limited feedback value. In addition, a small number of students perfunctorily participate in peer evaluation with arbitrary scoring and empty comments, undermining the scientificity and effectiveness of mutual evaluation activities.

5.3 Inherent Limitations of Intelligent Evaluation and Lack of Humanistic Assessment

AI writing evaluation tools excel in standardized superficial linguistic error correction, but cannot accurately interpret humanistic and critical elements such as thematic conception, emotional expression, logical depth and innovative thinking. Its judgment of contextual adaptability and personalized writing style has obvious deficiencies. Excessive reliance on AI evaluation tends to lead to templated and homogenized student writing, restricting the cultivation of innovative and personalized writing ability. This fully reflects the core principle that technology empowerment must be people-oriented in foreign language teaching.

5.4 Optimization Strategies

Targeting the above problems, three optimization paths are proposed based on practical teaching experience. First, carry out hierarchical special training on AI tool application. With positive and negative case comparison, guide students at different levels to master the skills of screening and optimizing AI feedback, and establish the learning concept of “AI as auxiliary tool and critical thinking as core” to eliminate technological dependence.

Second, refine peer evaluation standards. Construct a three-dimensional assessment

mechanism covering scoring results, comment quality and feedback adoption effect, and implement public display of evaluation results and teacher review system to standardize peer evaluation behaviors and narrow the quality gap of mutual assessment.

Third, adhere to the humanistic education orientation. Clarify the division of functions: AI is responsible for standardized linguistic optimization, while teachers and students undertake humanistic and critical evaluation. Balance technological empowerment and humanistic education to ensure the innovation and personalization of writing teaching.

6. Conclusion

Based on the policy background of new-era educational evaluation reform, this study focuses on the practical pain points of English writing teaching evaluation in undergraduate colleges. Integrating formative assessment theory and multi-subject evaluation theory, it constructs a four-stage progressive multi-subject collaborative evaluation model including student self-evaluation, peer evaluation, critical AI optimization, and teacher whole-process diagnosis. A semester-long teaching practice verifies the practical effectiveness of the model.

The research results confirm that the proposed evaluation system effectively solves the drawbacks of traditional writing teaching, including single evaluation dimension, delayed feedback, neglected learning process and low student initiative. It significantly improves students' writing performance, autonomous learning ability and metacognitive critical thinking, reshapes the classroom teaching ecology, and realizes the educational goal of promoting teaching and learning through scientific evaluation.

Compared with existing studies, this research realizes the refined, procedural and verified implementation of multi-subject collaborative evaluation. It clarifies the auxiliary

positioning and critical application path of AI tools, balances technological empowerment and humanistic education, and makes up for the deficiencies of insufficient innovation, weak empirical evidence and poor operability in current writing evaluation research. The universal practical paradigm can be extended to foreign language writing courses in various universities, with high replicability and popularization value. It provides operable and verifiable practical references for college foreign language education evaluation reform and intelligent writing teaching innovation.

Future research can expand the sample size, develop differentiated evaluation strategies for students with different learning levels, optimize the integration mechanism of intelligent evaluation and humanistic evaluation, and further improve the human-machine collaborative multi-subject evaluation system, so as to promote the high-quality and connotative development of college foreign language education in the new era.

References

- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74. <https://doi.org/10.1080/0969595980050102>
- Central Committee of the Communist Party of China & State Council. (2020). 深化新时代教育评价改革总体方案[*The overall plan for deepening the reform of educational evaluation in the new era*]. http://www.gov.cn/zhengce/2020-10/13/content_5551032.htm
- Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Sage Publications.
- Knoch, U. (2009). *Diagnostic writing assessment: The development and validation of a rating scale*. Peter Lang.
- Liu, N. F., & Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education*, 11(3), 279–290.

<https://doi.org/10.1080/13562510600680582>

- Qin, L. (2026). Validating and extending ecological affordance theory: An exploratory mix–method study comparing traditional and flipped EFL classrooms. *Journal of Language*, 2(1), 1–37. <https://doi.org/10.64699/26KJUI3673>
- Rutherford, S. et al. (2025), Assessment is learning: Developing a student–centred approach for assessment in higher education. *FEBS Open Bio*, 15(1), 21–34. <https://doi.org/10.1002/2211–5463.13921>
- Shaik, T. et al. (2022). A review of the trends and challenges in adopting NLP for education feedback analysis. *IEEE Access*, 10, 56720–56739. <https://doi.org/10.48550/arXiv.2301.08826>
- Stevenson, M., & Phakiti, A. (2014). The effects of computer–generated feedback on the quality of writing. *Assessing Writing*, 19, 51–65. <https://doi.org/10.1016/J.ASW.2013.11.007>
- Yue, Q. et al. (2026). Generative AI and the transformation of translation education: A systematic review of empirical studies. *Journal of Language*, 2(1), 184–226. <https://doi.org/10.64699/26MBAW1771>