Education and Social Work
ISSN Print: 3079-515X
ISSN Online: 3079-5168
https://doi.org/10.63313/ESW.9181

Formative Assessment in College English Writing Instruction: Effects on Revision Quality and Learner Autonomy

Chen Tang
Philippine Christian University, Manila 1004, Philippines
Hunan Institute of Science and Engineering, Yongzhou 425199, China
Email: 515083667@qq.com

How to cite this paper: Tang, C. (2026). Formative assessment in college English writing instruction: Effects on revision quality and learner autonomy. Education and Social Work, 6(1), 57–69. ISSN Print: 3079-515X, ISSN Online: 3079-5168.
https://doi.org/10.63313/ESW.9181
Published: 2026-11-09

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



Abstract

Writing remains one of the weakest productive skills among Chinese EFL learners, and teacher grading under a summative assessment regime often fails to translate into substantive revision on the part of students. This study employed a quasi-experimental, mixed-methods design to examine the effects of formative assessment on revision quality and learner autonomy in college English writing instruction. Over a 16-week semester, two parallel classes received either a formative assessment intervention (combining transparent assessment criteria, teacher formative feedback, peer review, and reflective journals) or conventional teacher grading. Results indicated that (1) the experimental class significantly outperformed the control class on the "global revision" dimensions of content improvement and discourse organization, whereas no significant between-group difference was found in "local" language corrections; (2) learners in the experimental class showed significant gains in autonomy, particularly in goal setting and self-monitoring; and (3) interview and journal data suggested that the internalization of assessment criteria and the "reader awareness" activated through peer review served as key mediators of autonomy growth. The study argues for an operational formative assessment cycle feasible in large-class EFL writing instruction, one that shifts students from "correcting errors" to "revising meaning."

Keywords

Formative Assessment; EFL Writing; Revision Quality; Learner Autonomy; Feedback Literacy

1. Introduction

Writing, as the most comprehensive productive language skill, has long been regarded as a major stumbling block in college English instruction in China. Despite the considerable time and effort teachers devote to grading compositions, student essays often follow a "one-draft" pattern: after graded papers are returned, students

either make only surface-level corrections or ignore the feedback altogether. Feedback thus fails to translate into genuine learning behavior. This inefficient cycle of "diligent teacher grading, indifferent student response" reflects the dominance of summative assessment in college EFL writing instruction and its inherent limitations.

Formative assessment advocates embedding evaluation within the learning process so that continuous feedback and adjustment serve learning itself (Black & Wiliam, 1998). The recent notion of feedback literacy further stresses that the value of feedback ultimately depends on students' capacity to understand, judge, and use feedback information (Carless & Boud, 2018). In writing instruction, whether formative assessment can actually improve the quality of students' revisions and cultivate their autonomy as writers remains an empirical question. Existing studies have largely addressed either feedback effectiveness or learner autonomy in isolation; integrated empirical research framing assessment type, revision quality, and learner autonomy within a single design is scarce, and evidence from the Chinese context is particularly limited.

Against this backdrop, the present study conducted a 16-week quasi-experiment in an authentic instructional setting to answer three research questions:

(1) On which dimensions does students' revision quality change significantly after the formative assessment intervention?

(2) What effects does formative assessment have on learners' autonomy?

(3) How do students perceive the role of formative assessment in shaping their revision behavior and autonomy?

2. Literature Review

2.1. Theoretical Foundations of Formative Assessment

Black and Wiliam's (1998) systematic review of over 250 studies established the pedagogical standing of formative assessment, showing effect sizes that exceed most educational interventions. Nicol and Macfarlane-Dick (2006) linked formative assessment to self-regulated learning and proposed seven principles of good feedback practice, emphasizing that feedback should help learners clarify goals, monitor performance, and regulate behavior. Hattie and Timperley (2007) demonstrated the conditions of feedback effectiveness through four levels of feedback information (task, process, self-regulation, and self). Together, this body of theory converges on a core proposition: the function of assessment lies not in judging but in providing actionable information to learners.

2.2. Feedback and Revision in L2 Writing

The relationship between feedback and revision is a central topic in L2 writing research. Ferris (1997) found an association between the specificity of teacher commentary and students' revision behavior, though teacher commentary was

unevenly distributed across types. In the Chinese college EFL context, Yang, Badger, and Yu (2006) compared teacher and peer feedback and found that the two differed in the types of errors each addressed, and that teacher feedback was not always students' first choice. Zhang and Hyland (2018) examined student engagement with teacher and automated feedback, characterizing engagement as a dynamic process involving cognitive, affective, and behavioral dimensions. Studies of Chinese classrooms by Han and Hyland (2015) and Zheng and Yu (2018) further indicate that learner uptake of written corrective feedback is shaped by the interaction of individual and contextual factors. Tian and Zhou (2020) revealed the dynamic and reciprocal ways learners engage with multiple feedback sources in online writing contexts.

Regarding revision quality, the literature commonly distinguishes local revisions (surface-level corrections of spelling and grammar) from global revisions (content restructuring and discourse reorganization). Prior evidence suggests that without an understanding of assessment criteria, students' revisions tend to remain superficial.

2.3. Learner Autonomy and Writing

Learner autonomy refers to learners' capacity to take charge of their own learning. In writing, autonomy manifests as the ability to independently evaluate one's own and others' texts and to apply criteria consciously when revising. Conventional teacher grading, however, objectively reinforces students' dependence on teacher authority, producing a "feedback-dependence" paradox: the more information students receive, the weaker their capacity for independent judgment becomes. By involving students in the assessment process (e.g., self-assessment with rubrics and peer review), formative assessment can theoretically "internalize" assessment criteria as writers' self-monitoring capacity (Nicol & Macfarlane-Dick, 2006), thereby fostering autonomy.

2.4. Research Gap

Taken together, existing studies provide a solid foundation but leave two gaps. First, most studies address either feedback effectiveness or autonomy alone; few quasi-experimental studies simultaneously examine the dual effects of formative assessment on revision quality and learner autonomy. Second, qualitative evidence (students' understanding of the assessment process) is rarely integrated with quantitative results. The present study seeks to fill these gaps.

3. Methodology

3.1. Research Context and Participants

The study was conducted in a 16-week College English course (2 class hours per week, 32 hours in total) at a provincial university in central China in the 2025

autumn semester. College English is a compulsory course for non-English majors, and writing is assessed within the final examination, which arguably reinforces summative orientation. Using cluster convenience sampling, two parallel classes taught by the same instructor (a female lecturer with 12 years of teaching experience and a background in EFL writing instruction) were selected: an experimental class of 45 students (19 males, 26 females, mean age 19.2) and a control class of 43 students (21 males, 22 females, mean age 19.4). Both classes were second-year non-English majors of comparable academic standing.

To ensure instructional equivalence beyond the intervention, the two classes used the same textbook, covered identical teaching content, completed the same four writing tasks under the same time limits, and received the same amount of total instructional time. The only systematic difference lay in the assessment and feedback procedure applied to the writing tasks. Table 1 presents the two classes' entrance English scores and all pretest measures.

Table 1. Participant Information and Pretest Homogeneity Tests

Measure	Experimental (n=45)	Control (n=43)	t	p
Gender (M/F)	19/26	21/22	—	—
NMET English score M(SD)	108.4(11.2)	109.6(12.0)	0.48	0.634
Writing pretest M(SD)	8.12(1.64)	8.21(1.58)	0.26	0.795
Autonomy pretest total M(SD)	86.53(9.47)	87.21(9.82)	0.33	0.742

As Table 1 shows, no significant between-class differences were found on any pretest measure ($p > 0.05$), satisfying the homogeneity assumption for quasi-experimental comparison. Written informed consent was obtained from all participants after the purpose of the study was explained; participation was voluntary, non-participation carried no academic consequences, and all data were anonymized (students were coded S01–S88). The study protocol was reviewed and approved by the teaching affairs office of the participating university.

3.2. Intervention

The control class received regular instruction: the teacher lectured on writing skills using the textbook's writing unit, students completed four argumentative writing tasks in class ("Should College Students Take Part-time Jobs?", "My View on Online Learning", "The Importance of Reading Classics", and "Is Praise Better than Criticism?"), and the teacher graded each essay with error corrections and a score (out of 15) before returning it; rewriting was not required.

Within the same tasks and class hours, the experimental class implemented a formative assessment cycle. Each writing task (one 4-week cycle, four cycles in total) comprised five stages:

(1) Making criteria transparent (about 30 minutes, Week 1 of each cycle). Before writing, students received and discussed an argumentative essay rubric

covering content, organization, and language, each dimension described at four levels (Excellent 4 / Good 3 / Pass 2 / Needs Improvement 1) with concrete descriptors and sample excerpts from anonymous student papers. The rubric was developed by the research team, reviewed and revised by two associate professors specializing in English writing and one native-speaker instructor (content validity review), and piloted in another parallel class of 44 students, after which two ambiguous descriptors were reworded based on student feedback.

(2) Teacher formative feedback (returned in Week 2). The teacher did not grade first drafts but provided descriptive comments on "improvable points" using a three-move pattern of "identifying a problem + probing question + suggestion" (e.g., "What is the relation between the example in Paragraph 2 and your claim? Could you add a specific statistic?"). Comments were 80–120 words per essay and avoided explicit error correction of grammar and spelling, which remained available to the control class through conventional grading. This design decision isolates the effect of formative feedback on global revision from that of error correction on surface revision.

(3) Peer review (about 25 minutes in class, Week 2). Working in fixed pairs re-rotated every cycle, students exchanged first drafts and completed a structured feedback sheet with reference to the rubric. The sheet required three entries: the essay's strongest aspect; the aspect most in need of improvement; and one concrete, actionable suggestion. Before the first peer review session, a 20-minute training session was conducted using a sample essay, during which the teacher modeled how to locate evidence in the rubric and phrase constructive comments.

(4) Revision (homework, within one week). Students integrated teacher feedback, peer feedback, and their own judgment to produce a second draft. Both drafts were submitted, which allowed revision quality to be operationalized as first-to-second-draft improvement.

(5) Reflective journal (homework). Students wrote a roughly 150-word Chinese reflection answering three questions: Which feedback did I adopt, and why? Which feedback did I reject, and why? How do I plan to improve my next draft?

Implementation fidelity. To protect internal validity, three fidelity measures were taken. First, the instructor followed a written implementation protocol specifying the timing, duration, and materials of each stage; weekly meetings between the instructor and the researchers checked adherence. Second, a fidelity checklist (12 items covering, e.g., "rubric distributed before writing," "no scores given on first drafts," "peer review completed in class") was completed for all four cycles; observed implementation rates were 100% for Stages 1–4 and 97% for Stage 5 (three students missed one journal each). Third, the control class was monitored to ensure no rubric or structured peer review leaked into its instruction; the instructor confirmed that her routine practice in the control class was identical to that of previous semesters.

The 16-week implementation schedule is summarized in Table 2.

Table 2. The 16-Week Implementation Schedule

Week	Experimental class	Control class
1	Pretest (writing + autonomy questionnaire); rubric training	Pretest (writing + autonomy questionnaire); regular instruction
2–5	Task 1: criteria → drafting → teacher feedback → peer review → revision → journal	Task 1: drafting → teacher grading
6–9	Task 2 (same cycle)	Task 2 (same grading)
10–13	Task 3 (same cycle)	Task 3 (same grading)
14–15	Task 4 (same cycle, compressed)	Task 4 (same grading)
16	Posttest (autonomy questionnaire); interviews; posttest writing	Posttest (autonomy questionnaire); posttest writing

Note: the writing pretest was a stand-alone argumentative task administered in Week 1 and excluded from analysis as an intervention task; the posttest writing was used only to report overall writing quality and is not analyzed in this paper, which focuses on revision quality.

3.3. Variables and Instruments

The independent variable was assessment type (formative assessment vs. conventional teacher grading); the dependent variables were revision quality and learner autonomy.

3.3.1 Revision Quality Rating Scale. Revision quality was defined as the quantity and quality of improvements from the first to the second draft, operationalized in three dimensions (Table 3), each scored 0–5 for a total of 0–15.

Table 3. The Revision Quality Rating Scale (excerpts)

Dimension	Operational definition	Typical 5-point performance	Typical 1-point performance
Content improvement	Deepened claims, added evidence, improved claim–evidence fit	New specific evidence (data, examples) highly relevant to the claim	Only isolated word substitutions or deletions
Organizational restructuring	Adjusted paragraph logic, added/deleted cohesive devices, improved opening–closing coherence	Paragraphs resequenced with transitional sentences added	No structural adjustment
Language correction	Number and accuracy of grammatical and lexical corrections	Systematic corrections without introducing new errors	Sporadic, random changes

The scale was developed through a three-step process: (a) a draft was written based on Faigley and Witte's classic taxonomy distinguishing surface changes from meaning-level changes and on the global/local revision distinction in the L2 writing literature; (b) two associate professors specializing in English writing reviewed the draft scale against the rubric dimensions to ensure construct alignment; (c) the scale was piloted on 20 first–second draft pairs from the pilot class, and descriptors of two dimensions were sharpened after raters reported ambiguity. In the main study, two English teachers not involved in the teaching independently and blindly rated all first–second draft pairs. Double ratings on a random 30% subsample yielded Cohen's Kappa of 0.82 (almost perfect agreement); remaining disagreements were adjudicated by a third rater, and rating disputes were resolved by consensus meetings after each scoring round.

3.3.2 Learner Autonomy Questionnaire. A widely used Chinese learner autonomy questionnaire (adapted version) was administered before and after the intervention, comprising 28 five-point Likert items across four subscales: goal setting (6 items), strategy selection (8 items), self-monitoring (7 items), and self-evaluation (7 items). Items were rated from 1 (strongly disagree) to 5 (strongly agree), with four negatively worded items reverse-scored. The full scale showed good reliability in the present sample (Cronbach's $\alpha = 0.89$; subscale αs = 0.78, 0.81, 0.83, 0.80). Exploratory factor analysis (principal axis factoring, promax rotation) confirmed the four-factor structure, accounting for 58.6% of cumulative variance.

3.3.3 Qualitative Data. After the intervention, 12 students from the experimental class were stratified by pretest performance (four from each of high, middle, and low bands) for semi-structured interviews of about 20 minutes each. The interview protocol centered on three questions: How did you understand the rubric and the feedback? What did peer review bring you? How do you evaluate your own revision behavior? Probing questions explored concrete episodes ("Can you give an example of a comment you adopted and what you did with it?"). In addition, all 180 reflective journals were collected. Interview transcripts and journals were imported into NVivo 12 for coding.

To strengthen trustworthiness, three strategies were employed: triangulation across data sources (interviews, journals, and rated texts), member checking (six interviewees verified the summaries of their interviews), and an audit trail documenting all coding decisions.

3.4. Data Analysis

Quantitative data were analyzed in SPSS 26.0 with the significance level set at $\alpha = 0.05$. The analytical procedure was: (1) independent-samples t tests to verify pretest homogeneity (Table 1), with Levene's and Shapiro–Wilk tests confirming homogeneity of variance and normality (all $p > 0.05$; Levene $F = 1.02–2.87$); (2) paired-samples t tests for within-group pre–post comparisons and independent-samples t tests for between-group posttest comparisons; (3) one-way ANCOVA with pretest scores as covariate to test between-group differences after controlling for initial levels, reporting partial η^2, and Cohen's d for within-group comparisons, interpreted following Cohen's benchmarks (0.2 small, 0.5 medium, 0.8 large). Missing data were minimal (six second drafts and three journals, 3.3% and 1.7% respectively) and were handled by available-case analysis; Little's MCAR test was non-significant ($p = 0.412$), supporting this treatment.

Two supplementary analyses were conducted. First, task-level trend analysis computed between-group effect sizes at each of the four tasks to trace the temporal unfolding of the intervention effect. Second, a feedback adoption analysis quantified, for each first draft, the proportion of teacher feedback points and peer feedback points that were visibly addressed in the second draft, based on the two raters'

agreement (Kappa = 0.79). Finally, Pearson correlations were computed between autonomy gain scores and revision quality scores in the experimental class to probe the hypothesized association between the two dependent variables.

Qualitative data were analyzed following Braun and Clarke's six-phase thematic analysis. Two researchers independently open-coded 30% of the data and negotiated an initial framework (41 first-order codes), which was subsequently condensed into 14 second-order codes and finally three themes; inter-coder agreement was 0.86, with discrepancies resolved through discussion. Throughout the analysis, deviant-case sampling was deliberately adopted to examine students whose autonomy scores did not improve, guarding against an overly positive account of the intervention.

4. Results

4.1. Revision Quality: Formative Assessment Enhanced Global Revision

A total of paired first–second drafts were collected: 180 in the experimental class (45 × 4) and 172 in the control class (43 × 4; 6 missing due to absence, 3.3% attrition). Table 4 presents revision quality scores by group.

Table 4. Descriptive Statistics for Revision Quality by Dimension, M(SD)

Dimension	Experimental (n=45)	Control (n=43)	t	p	Cohen's d
Content improvement	3.42(0.78)	2.31(0.71)	6.97	0.000*	1.47
Organizational restructuring	3.18(0.83)	2.26(0.74)	5.54	0.000*	1.16
Language correction	3.05(0.69)	2.87(0.72)	1.21	0.229	0.26
Total	9.65(1.87)	7.44(1.66)	5.85	0.000*	1.24

Note. Scores are means across the four writing tasks; *p < 0.001.

As Table 4 shows, the experimental class scored significantly higher than the control class on content improvement and organizational restructuring, with large effects (d = 1.47 and 1.16), as well as on the total score (d = 1.24). On language correction, however, the difference was not significant (p = 0.229, d = 0.26).

ANCOVA with pretest scores as covariate was conducted to control for the slight pretest difference (Table 5).

Table 5. ANCOVA Results for Total Revision Quality

Source	SS	df	MS	F	p	Partial η^2
Pretest (covariate)	38.62	1	38.62	12.73	0.001	0.13
Group (experimental/control)	152.08	1	152.08	21.36	0.000*	0.20
Error	262.41	86	3.05			

The ANCOVA revealed a significant group effect after controlling for pretest levels (F = 21.36, p < 0.001, partial η^2 = 0.20), with assessment type accounting for 20% of the between-group variance in revision quality.

Trend across tasks. Table 6 reports between-group effect sizes at each task. The

experimental class's advantage accumulated over time: differences were small at Tasks 1 and 2 (d = 0.31 and 0.52) but expanded markedly at Tasks 3 and 4 (d = 1.38 and 1.61). Within the experimental class, revision quality rose from M = 8.21 (SD = 1.71) at Task 1 to M = 10.62 (SD = 1.80) at Task 4, whereas the control class remained essentially flat (7.15 to 7.52). This pattern suggests that the effects of formative assessment require a 2–3 cycle "warm-up": students first needed to learn how to read and use the rubric before their revisions showed substantive gains, indicating a learning curve in the internalization of assessment criteria.

Table 6. Task-Level Means and Between-Group Effect Sizes for Total Revision Quality

Task	Experimental M(SD)	Control M(SD)	Cohen's d
Task 1	8.21(1.71)	7.15(1.62)	0.31
Task 2	9.08(1.75)	7.38(1.59)	0.52
Task 3	10.31(1.79)	7.46(1.68)	1.38
Task 4	10.62(1.80)	7.52(1.66)	1.61

Feedback adoption. The supplementary adoption analysis showed that students visibly addressed 68.4% of teacher feedback points and 53.7% of peer feedback points in their second drafts, and that adoption rates rose across cycles (teacher feedback: 58.2% → 76.5%; peer feedback: 41.9% → 63.3%). Notably, the adoption of peer feedback grew faster than that of teacher feedback, coinciding with the widening revision-quality gap after Task 2 and suggesting that peer review played an increasingly substantive role as students became more skilled reviewers.

The dimension-specific pattern merits attention: in both classes, language corrections were mainly driven by teacher-marked errors (73% of the experimental class's scores on this dimension came from correcting teacher-flagged errors), indicating that merely asking students to "revise on their own" does not necessarily improve linguistic accuracy; formative feedback and explicit language correction should complement each other.

4.2. Learner Autonomy: Significant Gains in Goal Setting and Self-Monitoring

Table 7 reports pre–post autonomy scores for both groups.

Table 7. Pre–Post Comparisons of Learner Autonomy by Group, M(SD)

Subscale	Group	Pretest	Posttest	t	p	Cohen's d
Goal setting	Experimental	18.24(3.11)	20.86(2.94)	5.21	0.000*	0.87
Control	18.35(3.20)	18.56(3.15)	0.38	0.705	0.07	
Strategy selection	Experimental	24.02(4.05)	24.88(3.96)	1.87	0.068	0.21
Control	24.31(4.12)	24.42(4.08)	0.15	0.882	0.03	
Self-monitoring	Experimental	20.86(3.42)	23.24(3.31)	5.03	0.000*	0.71
Control	21.14(3.55)	21.31(3.49)	0.29	0.774	0.05	
Self-evaluation	Experimental	23.41(3.68)	25.07(3.55)	3.92	0.000*	0.46
Control	23.28(3.74)	23.45(3.70)	0.27	0.789	0.05	
Total	Experimental	86.53(9.47)	94.05(9.12)	5.87	0.000*	0.88
Control	87.21(9.82)	87.74(9.65)	0.32	0.751	0.05	

Note. *p < 0.001.

The experimental class's posttest total score significantly exceeded its pretest score ($t = 5.87$, $p < 0.001$, $d = 0.88$), whereas the control class showed no significant change. At the subscale level, goal setting ($d = 0.87$) and self-monitoring ($d = 0.71$) showed the largest gains, followed by self-evaluation ($d = 0.46$); strategy selection did not reach significance ($p = 0.068$). The between-group posttest comparison (independent-samples t test) showed a significantly higher total score in the experimental class ($t = 4.92$, $p < 0.001$, $d = 0.75$), driven mainly by goal setting ($d = 0.72$) and self-monitoring ($d = 0.63$).

Association between autonomy gain and revision quality. In the experimental class, autonomy gain scores correlated positively and moderately with revision quality total scores ($r = 0.46$, $p < 0.01$), with the strongest association between the self-monitoring subscale and content improvement ($r = 0.51$, $p < 0.001$). Although the quasi-experimental design precludes causal claims at the individual level, this association is consistent with the theoretical account that autonomous self-regulation and substantive revision draw on the same underlying capacity—applying criteria to one's own text.

4.3. Qualitative Findings: Three Themes in the Growth of Autonomy

Thematic analysis of the interviews and journals yielded three themes and 14 second-order codes; the distribution of sources is presented in Table 8.

Table 8. Qualitative Themes and Code Frequencies

Theme	Second-order codes (examples)	Interview mentions (n=12)	Journal mentions (n=180)
Theme 1: "Seeing" the criteria brings epistemic awareness	Criteria made explicit; direction for revision; from "by feeling" to "by evidence"	11	96
Theme 2: Peer review activates "reader awareness"	Perspective taking; discovering shared problems; easier to review others than oneself	9	74
Theme 3: Shifting from "correcting errors" to "revising meaning"	Valuing content and logic; revision = rethinking; still relying on the teacher for right/wrong	7	52

Theme 1: "Seeing" the criteria brings epistemic awareness. Eleven of the 12 interviewees explicitly reported that the rubric changed their understanding of "a good essay." One intermediate-level student (S07) said in the interview: "I used to only know the teacher thought my argument was not fully developed, but I didn't know what 'fully developed' looked like. With the rubric, I know where I stand on each dimension." The journals contained 96 mentions of "revision now has a direction" or "no longer revising by feeling," the most frequent code, confirming the enabling role of transparent criteria in feedback understanding. Importantly, this awareness appeared to precede behavioral change: journals from Cycle 1 were dominated by criteria-related remarks, while concrete descriptions of revision actions became frequent only from Cycle 2 onward.

Theme 2: Peer review activates "reader awareness." Nine interviewees

identified peer review as a turning point in their revision behavior. Notably, several students reported that "finding faults in others' essays is easier than revising my own" (S03: "I can tell at a glance that a classmate's second paragraph is off topic, but I couldn't see the same problem in my own draft"). This "bystander effect" served as scaffolding for acquiring an evaluative perspective. Five students also mentioned the sense of responsibility peer review created: "He revised his essay based on my comments, so I took my reviewing more seriously next time." Journal data reinforced this account: 74 journal entries explicitly described applying to their own drafts the standards they had used on peers' drafts, suggesting that reviewing functioned as a rehearsal space for self-evaluation.

Theme 3: Shifting from "correcting errors" to "revising meaning," with individual differences. Roughly half of the interviewees (7/12) displayed a clear conceptual shift in their journals—from equating revision with "fixing grammar errors" early on to focusing on "making my ideas clear and fully developed" later. A representative trajectory came from S11, whose first journal read "I corrected nine grammar mistakes my teacher marked," but whose fourth journal read "I rewrote my second paragraph because the example didn't really support the point—that mattered more than the grammar." This shift typically occurred after the second or third cycle, consistent with the expanding between-group differences in revision quality at Tasks 3 and 4.

Nevertheless, individual differences were evident. Three students admitted they were "still not used to judging right and wrong by myself, and always want the teacher to mark the errors first," and their journals showed persistent error-fixation across all four cycles; these were exactly the students with low posttest autonomy scores. A further two students adopted a "compliance" orientation, reporting high adoption rates but minimal self-initiated changes beyond feedback points. These deviant cases suggest that for a minority of learners, 16 weeks of formative assessment functioned as an external procedure rather than an internalized practice, and that autonomy development may require longer scaffolding and more explicit strategy instruction for this subgroup.

5. Discussion

The results indicate that formative assessment's facilitative effect on revision quality was concentrated in global rather than surface-level revision. This is consistent with theoretical expectations: formative feedback delivers information at the task and process levels (Hattie & Timperley, 2007), guiding students to think about what to say and how to organize it, whereas language-level corrections rely heavily on the explicit identification of specific errors. The two types of feedback are not mutually exclusive. The finding suggests that expecting a single assessment approach to solve all writing problems is unrealistic; formative assessment and necessary language-focused correction should be complementary.

The significant gains in learner autonomy support Nicol and Macfarlane-Dick's (2006) core claim: when students are given opportunities to understand and apply assessment criteria, external evaluation is gradually transformed into self-evaluation and self-regulation. The qualitative data suggest that this transformation is mediated by the development of feedback literacy (Carless & Boud, 2018): students first "understand the criteria" (cognitive), then "apply the criteria" in peer review practice (behavioral), and finally "trust their own judgment" (affective and identity-related). At the same time, the non-significant gain in strategy selection and some students' continued reliance on authority remind us that cultivating autonomy is a long-term process; a 16-week intervention is insufficient to fundamentally reshape learners' beliefs about learning, echoing the constraining role of contextual factors in Chinese classrooms documented by Han and Hyland (2015).

Pedagogically, the study yielded an operational formative assessment cycle feasible in large-class settings. The key lies not in increasing teacher workload but in restructuring the feedback process: upfront rubrics lower the threshold for understanding feedback; peer review supplements feedback sources while training critical reading (Yang et al., 2006); and reflective journals make feedback processing visible. Students' participation in the feedback process is itself a form of writing instruction, confirming the "engagement" turn in feedback research (Zhang & Hyland, 2018; Zheng & Yu, 2018).

6. Conclusion

Through a 16-week quasi-experimental mixed-methods design, this study reached three conclusions. First, formative assessment significantly improved college EFL learners' global revision quality but had no significant effect on local, language-level corrections. Second, formative assessment effectively promoted learner autonomy, especially in goal setting and self-monitoring. Third, the internalization of assessment criteria and the "reader awareness" triggered by peer review constituted the key mechanisms of autonomy growth.

Several limitations should be acknowledged: the sample came from a single university, so generalization requires caution; a 16-week period remains short for testing long-term changes in autonomy; and autonomy was measured through self-report questionnaires, which may be subject to social desirability effects. Future research may conduct multi-site replication and longitudinal tracking, examine differential effects across proficiency levels, and explore the integration of formative assessment with intelligent writing feedback tools.

References

[1] 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

[2] Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81-112. https://doi.org/10.3102/003465430298487

[3] Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. Studies in Higher Education, 31(2), 199-218. https://doi.org/10.1080/03075070600572090

[4] Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. Assessment & Evaluation in Higher Education, 43(8), 1315-1325. https://doi.org/10.1080/03075079.2018.1442774

[5] Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. Language Teaching, 39(2), 83-101. https://doi.org/10.1017/S0261444806003399

[6] Ferris, D. R. (1997). The influence of teacher commentary on student revision. TESOL Quarterly, 31(2), 315-339. https://doi.org/10.2307/3588049

[7] Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. Journal of Second Language Writing, 15(3), 179-200. https://doi.org/10.1016/j.jslw.2006.05.002

[8] Zhang, Z., & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. Assessing Writing, 36, 90-102. https://doi.org/10.1016/j.asw.2018.02.004

[9] Zheng, Y., & Yu, S. (2018). Student engagement with teacher written corrective feedback in EFL writing: A case study of Chinese lower-proficiency students. Assessing Writing, 37, 13-24. https://doi.org/10.1016/j.asw.2018.03.001

[10] Tian, L., & Zhou, Y. (2020). Learner engagement with automated feedback, peer feedback and teacher feedback in an online EFL writing context. System, 91, 102247. https://doi.org/10.1016/j.system.2020.102247