Portfolio-Based Instruction and EFL Paragraph Writing: A Pretest–Posttest Comparison of Two University Classes

Prof. Nadhim Aldubai[1] , Prof. Omer Elsheikh Hago Elmahdi[2]

[1]Sanaa University, Affiliated to the Department of Languages & Translation, College of Arts & Human Science, Taibah University,

[2]The Open University of Sudan, Affiliated to the Department of Languages & Translation, College of Arts & Human Science, Taibah University, Saudi Arabia

DOI: https://doi.org/10.47772/IJRISS.2026.101000237

Received: 16 October 2026; Accepted: 21 October 2026; Published: 01 November 2026

ABSTRACT

This study examined the relationship between portfolio-based instruction and paragraph writing development among undergraduate learners of English as a foreign language at Taibah University, Saudi Arabia. Two intact civil engineering classes of 30 students each took part in an eight-week pretest–posttest comparison. One class (M29) kept its written assignments and instructor feedback in a cumulative portfolio that students could review when they wrote again; the other class (M32) followed the same routine of corrected assignments and redrafting but kept no portfolio. Every pretest and posttest paragraph was scored independently by two human raters on a 20-point analytic rubric covering content, organization, grammar, vocabulary, and mechanics, and each student's score was the mean of the two raters' totals. Agreement between the raters was high (ICC(2,1) = .96, 95% CI [.94, .97]; 98.3% of paired scores within one point). The portfolio class improved from 11.93 to 16.47 (mean gain 4.53 points), and the traditional class from 11.12 to 13.02 (mean gain 1.90 points). The difference in gains was 2.63 points (95% CI 1.97 to 3.30; Hedges' g = 2.03); it was 2.27 to 3.00 points when each rater's scores were analysed separately, and 2.68 points after adjustment for pretest scores. Because the portfolio condition was implemented in one class per condition and taught by one instructor, the results describe an association specific to these two classes and cannot be attributed to the portfolio arrangement alone. The study offers a transparent account of a small classroom comparison and a basis for replication across several classes.

Keywords: portfolio-based instruction; EFL paragraph writing; corrective feedback; formative assessment; university writing

INTRODUCTION

Paragraph Writing as a University Learning Task

Writing a paragraph asks learners to do several things at once. A writer must establish a focus, select relevant information, order supporting details, and express the relationships among ideas in sentences that a reader can follow. Grammatical accuracy and vocabulary matter, but they do not exhaust the task: a paragraph may contain individually acceptable sentences and still be hard to understand because its central idea is unclear or its support is poorly connected, while a coherent paragraph can communicate its meaning despite some language errors. Instruction and assessment, therefore, need to attend both to the organization of meaning and to the linguistic resources used to express it.

For university students who study a disciplinary subject through, or alongside, English, paragraph writing is a manageable setting in which to build these resources. A short text lets instructors respond to a complete piece of writing without requiring students to manage the demands of an extended report. In the civil engineering context of this study, the outcome of interest is foundational English paragraph writing, not the production of technical reports. The ability to develop an idea clearly is nevertheless an educationally important starting

point for the writing that students will later need in laboratory reports, project documentation, and professional correspondence.

From Feedback to Feedback Use

A recurrent instructional question concerns what happens after a marked assignment is returned. Corrective feedback can identify an error or suggest a stronger organization, but receiving a comment is different from understanding it and applying it in the next piece of writing. An instructor may spend considerable time correcting work without knowing whether students consult those corrections when they write again. A portfolio addresses this practical problem by keeping earlier texts and the responses to them available as a cumulative record. The record places a student's previous attempts, the instructor's comments, and the next task within reach of one another, so that learning can be carried forward from one assignment to the next instead of ending when each paper is returned.

The educational value of this arrangement depends on what students do with the record, and it is for this reason that a portfolio is best understood as an instructional practice and not merely a storage device. When students and teachers use the portfolio to look back, notice recurring difficulties, and plan a response, it becomes a working tool for development.

The Present Study

This study compares two classes that both received writing instruction and corrective feedback. The comparison therefore concerns the systematic retention and subsequent use of assignments and feedback, which was a feature of the portfolio class (M29), set against regular instruction with corrective feedback in the comparison class (M32). Framed this way, the comparison directs attention to continuity across tasks and asks what a cumulative portfolio arrangement adds to instruction that already includes feedback. It treats M32 as a class that was learning in the ordinary way, with the benefit of teaching, examples, practice, and correction.

Each class contributed 30 matched pretest and posttest scores. Paragraphs were scored independently by two raters on a 20-point analytic rubric of five equally weighted dimensions, and the study examines overall change within each class, the difference in mean change between classes, the distribution of individual outcomes, and the consistency of the two raters. The study concerns overall paragraph writing performance as captured by the rubric total. Because there is one class in each condition and one instructor, the instructional condition is confounded with every other difference between the classes, and the comparison is reported as an association within these two classes and not as an estimate of the effect of portfolios. Effects on writing anxiety, motivation, or each separate rubric dimension, and comparisons between digital and paper portfolios, are matters for other studies with measures designed for them.

Research Questions

The study addresses four questions:

1. How did total paragraph writing scores change from pretest to posttest within the portfolio class and within the traditional class?
2. How much did the two classes differ in their mean gains?
3. How widespread was improvement at the individual level, and to what extent did the scores of the two classes overlap?
4. How consistently did two independent raters score the paragraphs, and do the conclusions hold when each rater's scores are analysed separately and when pretest scores are taken into account?

The first question is answered by within-class summaries, the second by a direct comparison of gains, and the third by the distribution of individual gains and a plot of each student's two scores. The fourth is answered by agreement statistics for the two raters and by repeating the main comparisons on each rater's scores and with adjustment for the pretest. Together they give a fuller account than a single statement about statistical significance.

LITERATURE REVIEW AND CONCEPTUAL FRAMEWORK

Portfolios as records of Writing Development

A writing portfolio can serve several purposes. It may document selected accomplishments for final evaluation, preserve a sequence of working drafts, or support discussion of development over time. These purposes are worth distinguishing, because a collection of polished final pieces provides different evidence from a record that includes initial attempts, feedback, and revisions. Hamp-Lyons and Condon (2000) offer a comprehensive treatment of portfolio-based writing assessment, including the theory behind portfolios and the effects they can have on a writing programme. Lam (2017) reviews writing portfolio scholarship with attention to the relationship between research and classroom implementation, and Wang and He (2020) review research on portfolios in ESL and EFL contexts. Together these works support treating portfolio work as part of writing pedagogy rather than as an administrative addition at the end of a course.

The principle most relevant to the present study is continuity. An individual assignment shows what a learner produced on one occasion, whereas a sequence of retained assignments allows comparison across occasions. A student can consult a previous correction when a related problem occurs again, and an instructor can see whether the same problem persists. In this way a portfolio turns a series of separate tasks into a connected learning sequence. Its developmental value is greatest when students understand why drafts are retained and what they are expected to learn from them: when early attempts are treated as part of learning and not only as material for grading, students can experiment, revisit weaker drafts, and use the record to plan what to do next.

Formative Assessment and the Use of Evidence

Black and Wiliam (1998) place classroom assessment within the relationship between evidence of learning and subsequent teaching or learning action. Their account is relevant because it shows how a marked paragraph can function as more than a final judgment: it can indicate what should be addressed in the next attempt. The formative value of a score or comment lies in the use made of it, and a score that is recorded and then set aside plays a different instructional role from one that guides a later decision.

Applied to paragraph writing, the decision may concern the relevance of supporting details, the ordering of ideas, or a recurring grammatical pattern. The instructor can identify the problem, and the student then needs an opportunity to work on it. A portfolio can make that opportunity easier to organize by placing the earlier text beside the later attempt. The pedagogical proposal is temporal: information from one task should inform action in another.

The same reasoning explains why a traditional class can be expected to improve substantially. Students learn from classroom explanations, examples, exercises, and teacher correction even without a formal portfolio. A comparison between a portfolio class and a traditional class is therefore a comparison between two active learning environments, and the question of interest is whether the cumulative arrangement adds value to learning that is already taking place.

Feedback and Revision in Second Language Writing

Feedback on second language writing has been a central and at times contested topic. Truscott (1996) argued that grammar correction has no place in L2 writing classes, and Ferris (1999) responded with a case for its value, an exchange that made written corrective feedback a much-discussed topic in the field. Later syntheses give a more encouraging picture. Kang and Han (2015) combined 21 primary studies in a meta-analysis and found that written corrective feedback can improve the grammatical accuracy of second language writing, and Ferris (2010) discusses how L2 writing research and second language acquisition research on written corrective feedback can inform classroom practice. Hyland and Hyland (2006) describe the range of practices through which learners receive responses to their texts, and their account indicates that feedback is best described by its role and use, not merely by its presence. That a teacher corrected an assignment does not itself specify what was corrected, how the response was interpreted, or whether the learner applied it independently later.

Research in Saudi universities shows how feedback is provided and received in this setting. Alshahrani and Storch (2014) studied three teachers in a large Saudi university with strict guidelines on written corrective feedback and found that the teachers gave comprehensive indirect feedback, that most of it concerned mechanics, and that students preferred direct feedback, mainly on grammar. More recent studies continue to examine the match between what Saudi university teachers provide and what students prefer (Abdel Latif et al., 2024) and the effects of scaffolded feedback on Saudi students' writing accuracy over time (Alshahrani & Storch, 2025). This work points to the value of arrangements that help students use the feedback they receive, which is what a portfolio is designed to do.

Revision can involve changes of different depth. Replacing an incorrect verb form responds to a local problem, whereas adding a missing explanation or reorganizing the support for a claim changes how meaning is developed. Both kinds of change are useful, and an overall writing score can rise through either one or through a combination. A further distinction separates the improvement of a particular text from transfer to a new writing occasion. A student may produce a better draft by following detailed instructions, and the aim of a portfolio-based course is that the student also takes away a strategy that can be used independently on the next task. Assessments conducted on new tasks, such as the posttest in the present study, are therefore especially informative about development.

Self-Regulation, Feedback Literacy, and Engagement with Feedback

Nicol and Macfarlane-Dick (2006) connect formative assessment and feedback with students' capacity to regulate their own learning. Their framework is relevant because a portfolio makes prior performance available for review and comparison, and it can therefore support the cycle in which students recognize a difficulty, choose an action, and evaluate the result. When an instructional sequence repeatedly asks students to carry out these activities, they have the chance to become better at them.

Carless and Boud (2018) develop the concept of student feedback literacy, directing attention to the understandings and capacities that students need in order to use feedback productively. Research on engagement with feedback points in the same direction. Liu and Storch (2023) examined how learners in a university EAP class responded to teacher feedback on all aspects of their writing, comparing initial and revised texts. The learners took up almost all feedback suggestions, whatever their form or focus, yet their engagement with the feedback differed from one learner to another. This suggests that revision opportunities are taken up readily and that the quality of engagement is where learners differ, which is a reason to give students a structured place, such as a portfolio, in which to work with feedback.

An example illustrates the distinction. Suppose a teacher comments that a paragraph needs more support. A learner might add an unrelated sentence simply to make the paragraph longer, or might explain the central claim with a relevant reason or example. Placing the original comment beside the revision shows which response occurred and lets the instructor discuss the purpose of support with the student. A portfolio makes this kind of side-by-side comparison routine.

Empirical Studies of Portfolio-Based Writing Instruction

Studies in Saudi and wider Arab EFL contexts

Portfolio-based writing instruction has been examined in several Saudi and Arab classrooms. Alam and Aktar (2019) introduced formative portfolio assessment into a preparatory intensive English course in Saudi Arabia and used interviews to explore how students experienced it. They concluded that the approach encouraged autonomous learning by giving students more control over their own learning and assessment. Abduljawad (2024) worked with 120 male students in six classes at Yanbu Industrial College over 12 weeks, comparing paper-based and electronic portfolios, each combined with additional pedagogical strategies, against a control group. Both portfolio groups showed highly significant pre-to-post gains, whereas the control group's change was small, and paper-based and electronic portfolios did not differ significantly. Interviews in that study also indicated that heavy correction could leave students reluctant to read their errors, which supports concentrating feedback on the most important issues.

Studies elsewhere in the Arab world add further evidence. Obeiah and Bataineh (2016) compared a portfolio group with a conventionally instructed group of 20 Jordanian tenth-grade girls each, following the portfolio approach of Hamp-Lyons and Condon (2000), and reported that the portfolio group outperformed the control group in overall writing performance and in focus, development, organization, conventions, and word choice. Boumediene et al. (2016) examined a portfolio assessment model with third-year secondary students in Algeria using pretests and posttests and an analytic scoring rubric, and reported notable improvement in writing performance and in students' use of writing processes. These studies come mostly from school and college settings, and they leave room for evidence from university classes with a defined disciplinary focus.

Studies in other EFL contexts

Research from other EFL settings points in a similar direction. Fathi et al. (2020) randomly assigned 41 EFL learners to a portfolio-based instruction group or a control group that received regular instruction with no archiving of drafts, and reported through ANCOVA that the portfolio group improved more in writing performance and also showed a significant reduction in writing anxiety. Gebrekidan and Zeru (2023) conducted an eight-week quasi-experiment with 60 first-year students at Wollo University, comparing portfolio-based assessment with conventional assessment, and found significant differences in both writing performance and conceptions of writing. Pourdana and Tavassoli (2022) examined electronic portfolio assessment in which writing tasks received weekly teacher feedback, reflection logs, and revision, and reported significant progress in lower-level skills and moderate progress in higher-level skills in both groups of learners they studied. Farahian and Avarzamani (2018) examined the impact of portfolios on EFL learners' metacognition as well as their writing performance.

The studies differ in medium, genre, learner level, and the practices combined with the portfolio, such as reflection, peer assessment, and electronic tools. Taken together they show that portfolio practice has been examined in many contexts with encouraging results and that the portfolio condition often brings several practices at once.

Where the Present Study Fits

The present study adds evidence on a narrower question. It was carried out in Saudi university classes of civil engineering students studying paragraph writing. Both classes followed the same eight-week routine, in which every assignment was corrected, redrafted by students, and corrected again, and the instructional difference was whether assignments and feedback were kept together in a cumulative portfolio. This narrows the contrast to cumulative retention within a course that already includes corrective feedback and redrafting, although, as Section 7 explains, one class per condition cannot separate that contrast from other differences between the classes. The study complements studies in which the portfolio condition included additional practices.

Conceptual framework

The framework guiding the study begins with access to earlier work and feedback. Access allows review, review supports recognition of recurring problems, and deliberate responses to those problems contribute to improved later writing. Each step involves learner activity and is shaped by the clarity of instruction. The present study examines the outcome end of this pathway, namely performance on two paragraph writing assessments, and it asks whether the class that worked within a cumulative portfolio arrangement showed greater improvement than the class that did not.

METHOD

Research Design and Setting

The study used a quasi-experimental pretest–posttest design in an undergraduate civil engineering EFL writing course at Taibah University, Saudi Arabia. Two existing classes took part: M29, the portfolio class, and M32, the traditional comparison class. Because the classes were intact, students were not individually assigned to

conditions, and the instructional condition was implemented at class level. Scores were recorded for each student, and the analysis follows each student's own change from pretest to posttest.

3.2 Participants

The analytic sample consists of 30 students in M29 and 30 in M32, each with a pretest and a posttest score within the 20-point maximum. All 60 records are complete, so no scores were imputed. The students had completed a preparatory English programme before starting the writing course. Student names were removed from the analytic file, and figures display scores without identifying labels.

3.3 Instructional conditions

Both classes received instruction in paragraph writing and corrective feedback on their assignments. In M29, the paragraph assignments, instructor corrections, and feedback were retained in a portfolio for later review. The purpose of the portfolio was to connect successive writing experiences, so that each assignment could be read in relation to the ones before and after it. M32 followed traditional instruction: students received the same kind of teaching and corrective feedback, and a cumulative portfolio record was not maintained. The instructional contrast is therefore the cumulative portfolio arrangement, and feedback was part of both conditions.

Feature	M29 (portfolio)	M32 (traditional)
Students with complete pretest and posttest	30	30
Course length and meetings	Eight weeks, two per week	Eight weeks, two per week
Writing instruction	Provided	Provided
Corrected assignment and redraft at each meeting	Yes	Yes
Cumulative portfolio of assignments and feedback	Kept	Not kept
Allocation	Intact class	Intact class

The course ran for eight weeks in both classes, with two meetings per week of 75 minutes each, giving 16 meetings in all. At each meeting, one paragraph assignment was corrected by the instructor; students then prepared a revised draft, and the revision was corrected again at the following meeting. This routine was identical in the two classes. The only difference between the conditions was that M29 kept the assignments, corrections, and revised drafts together in a portfolio throughout the course, whereas M32 did not keep a portfolio. Instruction in both classes was based on the course textbook, Oshima and Hogue (2006), following a pacing schedule of Units 1 to 8. Some units, such as the unit on citations and documentation styles, were reviewed only briefly, and from each unit the instructor selected the material that strengthened the teaching of the paragraph. Teaching focused on paragraph structure and its components (the topic sentence, thesis statement, supporting details, body, and conclusion), with extensive exercises. The author taught both classes.

Assessment Tasks and Conditions

Writing was assessed on a 20-point analytic rubric of five equally weighted dimensions: content, organization, grammar, vocabulary, and mechanics, each worth up to four points. Content and organization concern the relevance and arrangement of ideas; grammar and vocabulary concern the linguistic resources used to express them; mechanics concerns conventions such as punctuation and spelling. The same rubric and scale were used in both classes at both testing occasions.

The pretest was given at the start of the course as a diagnostic test. Students were told that it would show where they stood and would help the instructor identify their weaknesses. The posttest was the midterm examination, held after the eight weeks of instruction. Students were not told that the assessments might later be used for research, so both tests were approached as ordinary course assessments. Each test lasted 45 minutes, and students wrote without a dictionary, notes, or earlier drafts, relying on what they could produce unaided.

The two tasks differed in topic, genre, and length requirement (Appendix B). The pretest asked for a descriptive and explanatory paragraph of 80 to 100 words entitled “My University Major”, with five guiding questions (what the major is, why it was chosen, what the student likes about it, what is difficult about it, and what job the student would like in the future). The posttest asked for a cause-and-effect paragraph of 100 to 120 words on “The effects of eating too much fast food”, explaining two or three effects on a person’s health and daily life, and it required suitable transition signals. Both tests asked for a topic sentence, supporting details, and a concluding sentence, and both were scored with the same rubric. Both classes sat exactly these two tests, at the same times and under the same conditions. Because the pretest was diagnostic and the posttest was a midterm examination, and because the prompts differed in topic, genre, scaffolding, and word limit, the two occasions are not equated in difficulty, and no procedure was used to equate them. Because both classes sat the same assessments under the same conditions, these differences are shared by the classes and bear less on the comparison of gains than on the absolute size of the gains, but a class-by-test interaction cannot be excluded.

Scoring Procedure

Each paragraph was scored independently by two human raters using the rubric: the first author (Rater 1), who taught both classes, and the second author (Rater 2). Each rater assigned a total out of 20 to every pretest and posttest paragraph, which gives 120 pairs of ratings (60 students at two occasions). A student's score at each occasion is the arithmetic mean of the two raters' totals, so scores can take half-point values. No score was rounded, replaced, or adjudicated, and no scores from any other source were used. The two raters' marks were recorded separately and are analysed here in their original form; they are reported for every student in Appendix A. The scripts given to Rater 2 were anonymised, and Rater 2 did not see Rater 1's marks at any stage. Rater 1 was the instructor of both classes, so expectancy effects on his marks cannot be ruled out; for this reason the main comparisons are repeated for each rater separately (Section 4.7).

Data Analysis

Gain was defined as posttest minus pretest, so positive values represent improvement, zero represents an unchanged total, and negative values represent a lower posttest score. For each class, means and sample standard deviations describe the pretest, the posttest, and the gain, and medians and observed ranges give complementary summaries of the gain distribution. The primary between-class comparison is the difference in mean gain, which is reported on the original 20-point scale because it is the most directly interpretable outcome.

Agreement between the raters was assessed on the 120 pairs of ratings with the intraclass correlation coefficient for absolute agreement between single raters, ICC(2,1) (Shrout & Fleiss, 1979; Koo & Li, 2016), with a 95% confidence interval based on the F distribution; with quadratic-weighted kappa (Fleiss & Cohen, 1973); with the percentage of pairs differing by 0, at most 1, and at most 2 points; and with the mean difference and 95% limits of agreement (Bland & Altman, 1986). Paired-samples t tests summarize change within each class, and Welch tests compare the classes at pretest, at posttest, and on gain scores; all tests are two-sided. Within-class standardized change is reported as d_z, the mean gain divided by the standard deviation of individual gains. The standardized difference between classes is reported as Hedges' g, computed with the pooled within-class standard deviation and the small-sample correction. The two standardized measures have different denominators and are interpreted separately (Lakens, 2013).

Three further analyses check the robustness of the main comparison. First, Wilcoxon signed-rank and Mann–Whitney tests were run on gains because the gain distribution in M32 departs from normality (Shapiro–Wilk $p = .011$). Second, posttest scores were regressed on class and pretest score (analysis of covariance) to allow for the small difference between the classes at the start. Third, the comparisons were repeated on each rater's scores separately. All inferential statistics treat the 60 students as independent observations. Because each condition is represented by a single class, between-class variation cannot be estimated, and the p values and confidence intervals describe the sample of students observed and do not support generalization to other classes. Figures were produced from the recorded scores; in the scatter plots, the area of a marker increases with the number of students at that coordinate, and numbers beside markers give the count.

RESULTS

Agreement between the raters

Table 1 summarizes agreement between the two raters on the 120 paragraphs. Agreement was high: ICC(2,1) = .96, 95% CI [.94, .97], with quadratic-weighted kappa = .96 and Pearson r = .96. The raters assigned identical totals to 16.7% of the paragraphs (20 of 120), totals within one point to 98.3%, and totals within two points to all of them; the largest discrepancy was two points. Exact agreement is modest on a 21-point scale, as is usual for holistic marks, whereas agreement within one point is almost complete. On average Rater 2 scored 0.07 points higher than Rater 1 (SD of differences = 0.94; 95% limits of agreement −1.78 to 1.91; Figure 1).

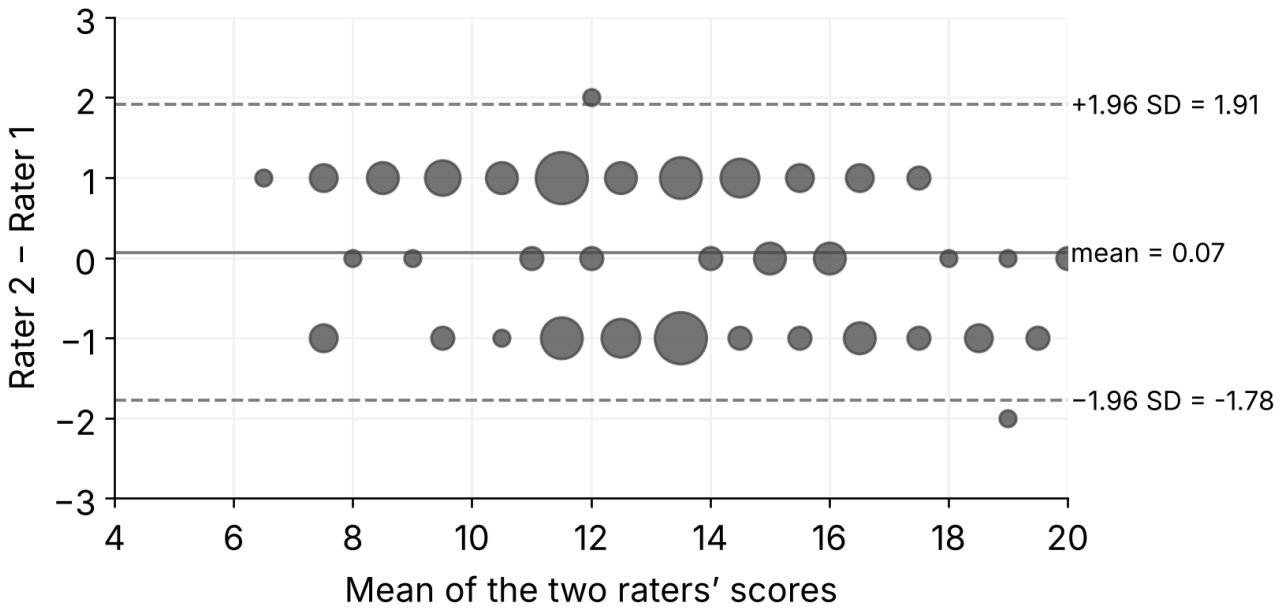
The raters differed systematically in one of the four class-by-occasion cells: Rater 2 gave M32 posttest paragraphs 0.50 points more than Rater 1 did, t(29) = 3.34, p = .002. Differences in the other three cells were small and not statistically significant (Table 1). Agreement on individual gains, which are differences between two imperfect measurements, was lower than agreement on scores (ICC(2,1) = .78), although the raters agreed on whether a student improved, stayed the same, or declined for 95% of students. Differences of a point or less in an individual student's change should therefore not be interpreted.

Table 1. Agreement between Rater 1 and Rater 2

Scope	n	Mean difference (R2 − R1)	SD of differences	Exact agreement (%)	ICC(2,1)	Paired t, p
All paragraphs	120	0.07	0.94	16.7	.96 [.94, .97]	.439
M29 pretest	30	0.00	0.91	20.0	.93	1.000
M29 posttest	30	0.00	0.95	23.3	.94	1.000
M32 pretest	30	−0.23	0.97	3.3	.90	.199
M32 posttest	30	0.50	0.82	20.0	.93	.002

Note. R2 − R1 is Rater 2's total minus Rater 1's total. ICC(2,1) is the two-way random-effects, absolute-agreement coefficient for a single rater; the interval for all paragraphs is a 95% confidence interval. p values are from paired t tests on the 30 (or 120) pairs.

Figure 1. Agreement between the two raters (Bland–Altman plot)



Note. Each marker is a paragraph (n = 120); marker area increases with the number of paragraphs at that point. The solid line is the mean difference and the dashed lines are the 95% limits of agreement.

Initial Scores

The two classes began at similar levels. The portfolio class had a pretest mean of 11.93 (SD = 2.41) and the traditional class a pretest mean of 11.12 (SD = 2.12), a difference of 0.82 points in favour of M29. The difference was not statistically significant, Welch t(57.08) = 1.39, p = .170, 95% CI [−0.36, 1.99], but it corresponds to about a third of a standard deviation (Hedges' g = 0.35), which is not negligible with 30 students per class. The analysis in Section 4.7 therefore also adjusts for the pretest.

Overall Scores and Gains

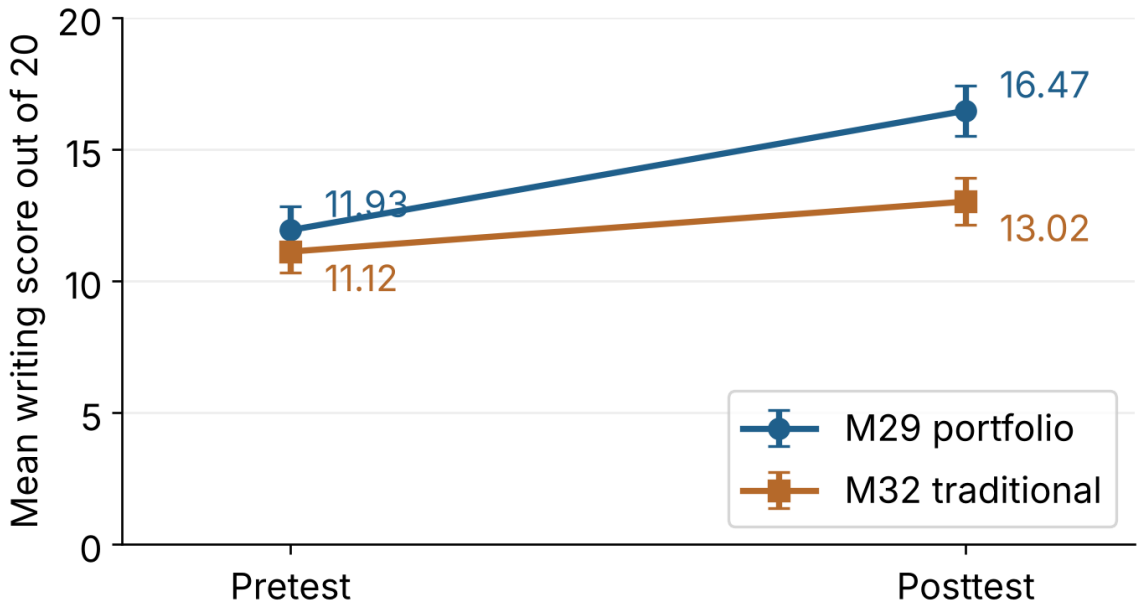
Both classes improved, and the portfolio class improved about 2.4 times as much on average (Table 2). M29 rose from a pretest mean of 11.93 to a posttest mean of 16.47, a mean gain of 4.53 points, and M32 rose from 11.12 to 13.02, a mean gain of 1.90 points. The classes started 0.82 points apart and finished 3.45 points apart, so the final gap reflects mainly the difference in improvement and only partly the difference in starting level.

Table 2. Descriptive statistics for writing scores (out of 20)

Class	n	Pretest M (SD)	Posttest M (SD)	Gain M (SD)	Median gain (range)	Improved / unchanged / declined
M29 portfolio	30	11.93 (2.41)	16.47 (2.59)	4.53 (1.66)	4.5 (1.5 to 8.0)	30 / 0 / 0
M32 traditional	30	11.12 (2.12)	13.02 (2.40)	1.90 (0.71)	2.0 (0.5 to 3.0)	30 / 0 / 0

Note. Scores are the mean of the two raters' totals. Gain = posttest − pretest. SD is the sample standard deviation.

Figure 2. Mean pretest and posttest scores by class



Note. Error bars are conventional student-level 95% confidence intervals, unadjusted for class clustering. Lines connect the two measured occasions; scores are out of 20 and are the mean of two raters.

Figure 2 shows the two observed means for each class. The steeper line for M29 reflects its larger average increase, and the rising line for M32 shows that students in the traditional class also improved. The figure displays a difference in degree between the classes: students in both classes made progress, and those in the portfolio class made more.

4.4 Changes within each class

Change from pretest to posttest was clear in both classes (Table 3). In M29, the mean gain of 4.53 points (SD = 1.66) was associated with $t(29) = 14.95$, $p < .001$, and a 95% confidence interval from 3.91 to 5.15; the standardized change was $d_z = 2.73$. In M32, the mean gain of 1.90 points (SD = 0.71) was associated with $t(29) = 14.62$, $p < .001$, a 95% confidence interval from 1.63 to 2.17, and $d_z = 2.67$. The standardized changes are similar and large in the two classes because gains within each class varied little relative to their mean; the raw gains, expressed in rubric points, are the more direct indicator of the educational scale of the change. Wilcoxon signed-rank tests led to the same conclusion in both classes ($p < .001$).

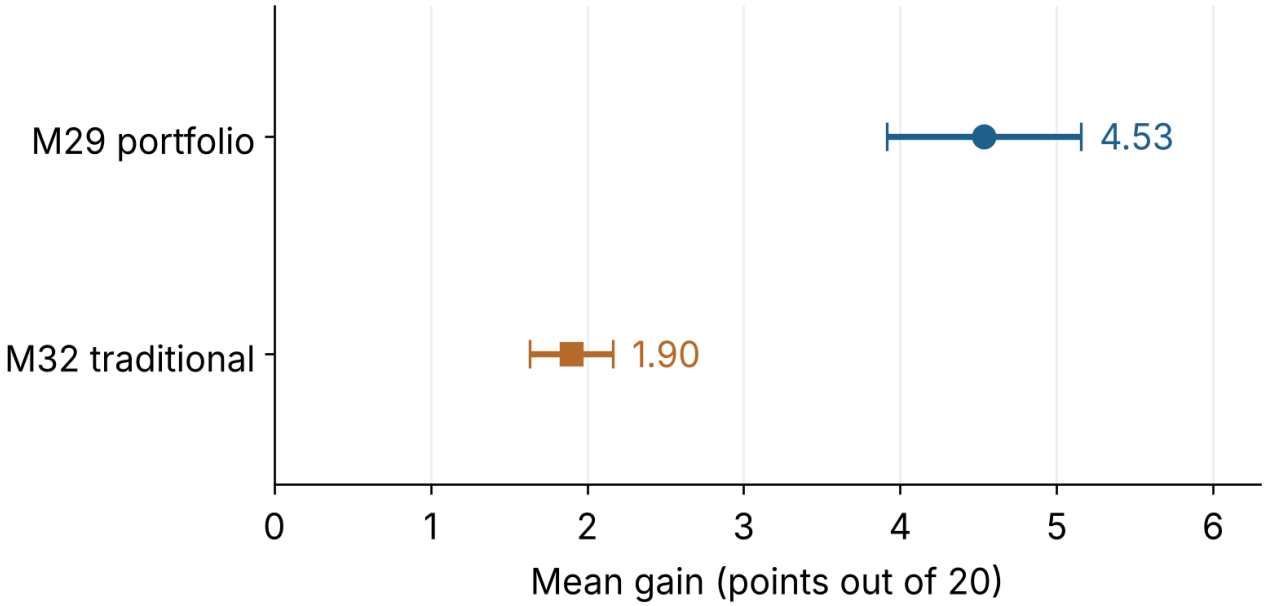
Table 3. Paired comparisons within each class

Class	Mean gain	95% CI	t(29)	p	d_z
M29 portfolio	4.53	3.91 to 5.15	14.95	< .001	2.73
M32 traditional	1.90	1.63 to 2.17	14.62	< .001	2.67

Note. d_z = mean gain divided by the standard deviation of gains.

Figure 3 compares the mean gains on the original score scale, with 95% confidence intervals for each class.

Figure 3. Mean gains by class



Note. Gain = posttest − pretest. Error bars are conventional student-level 95% confidence intervals, unadjusted for class clustering; n = 30 per class.

Difference between the classes

The comparison of gains answers the second research question directly. The portfolio class gained 2.63 points more than the traditional class (M29 minus M32), 95% CI [1.97, 3.30], Welch $t(39.31) = 7.98$, $p < .001$, with a standardized difference of Hedges' $g = 2.03$; a Mann–Whitney test agreed ($p < .001$). A difference of 2.63 points equals 13% of the 20-point scale. The posttest means differed by 3.45 points, 95% CI [2.16, 4.74], Welch $t(57.67) = 5.36$, $p < .001$, $g = 1.37$. The posttest difference equals the baseline difference of 0.82 plus the gain difference of 2.63, which provides a consistency check across the summaries in Table 4. These tests and intervals treat the students as independent; with one class per condition they describe the observed students and do not show that the result would recur in other classes.

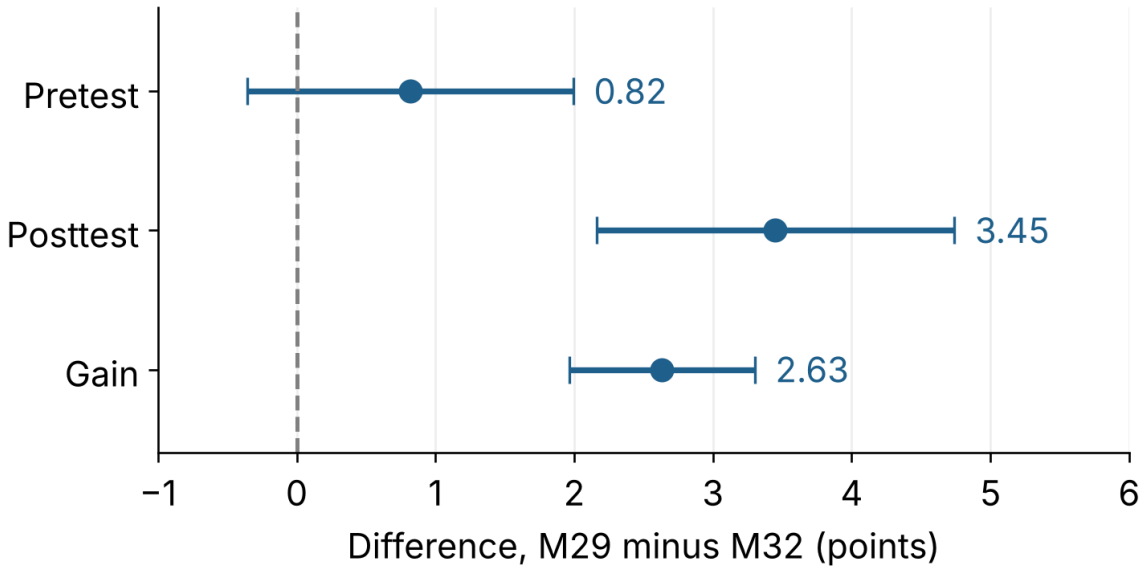
Table 4. Welch comparisons between classes (M29 minus M32)

Contrast	Difference	95% CI	t (df)	p
Pretest	0.82	−0.36 to 1.99	1.39 (57.08)	.170
Posttest	3.45	2.16 to 4.74	5.36 (57.67)	< .001
Gain	2.63	1.97 to 3.30	7.98 (39.31)	< .001

Note. Scores are the mean of the two raters' totals. Intervals and p values are not adjusted for class clustering.

Figure 4 summarizes the pretest, posttest, and gain differences between classes on the same score scale.

Figure 4. Between-class mean differences



Note. Dots show M29 minus M32; horizontal bars show conventional Welch 95% confidence intervals. The dashed line marks zero difference. Intervals are not adjusted for class clustering.

Individual patterns

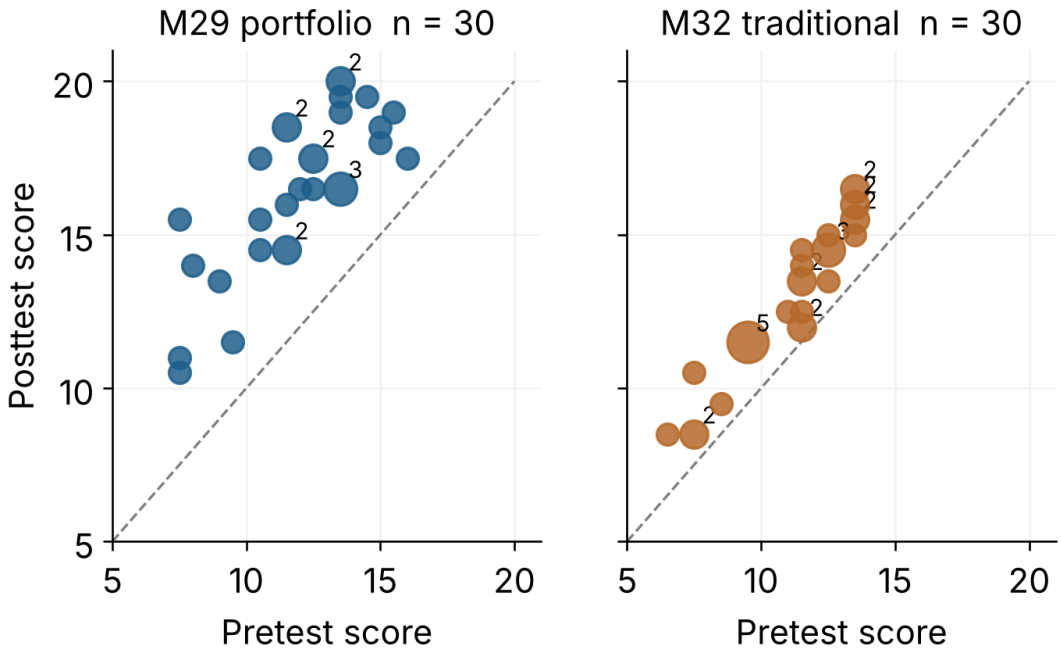
Every student in both classes improved on the two-rater mean (Table 5). In M29, gains ranged from 1.5 to 8.0 points with a median of 4.5; in M32, gains ranged from 0.5 to 3.0 with a median of 2.0. In M29, 18 students (60.0%) gained at least four points, and no student in M32 did; 28 students in M29 (93.3%) gained at least three points, against 4 in M32 (13.3%). Gains below one point, of which there were two in M32, are within the range of disagreement between the raters and should not be read as individual change. The scatter plots in Figure 5 show each student's pretest and posttest scores: all points in both classes lie above the diagonal that marks no change, and the M29 points lie further from it. The classes overlap in their posttest scores, so the higher class mean reflects a general upward shift and not a separation of every student in one class from every student in the other.

Table 5. Number of students at each gain value

Gain (points)	0.5–1.0	1.5–2.0	2.5–3.0	3.5–4.0	4.5–5.0	5.5–6.0	6.5–7.0	7.5–8.0
M29 portfolio	0	2	7	5	7	3	5	1
M32 traditional	7	15	8	0	0	0	0	0

Note. Gain = posttest − pretest on the two-rater mean, grouped in one-point bands (for example, 0.5–1.0 includes gains of 0.5 and 1.0). No student gained 0 or less.

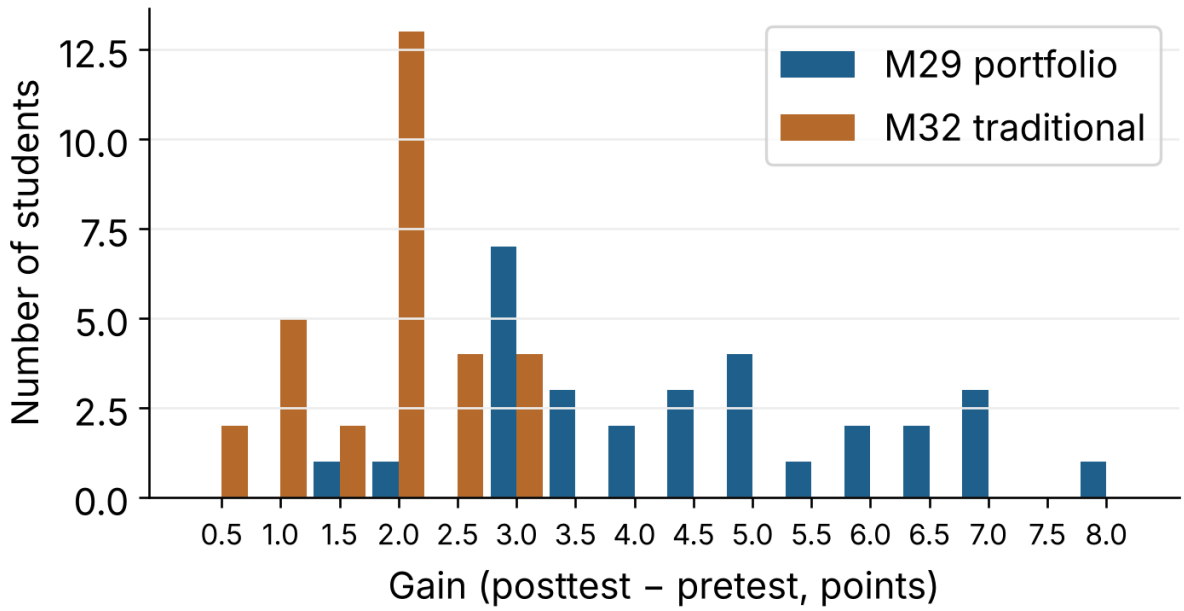
Figure 5. Individual pretest and posttest scores



Note. The dashed diagonal marks no change. Marker area increases with the number of students at a coordinate, and numbers beside markers give the count. All 60 students are represented; no student names are shown. Two students in M29 (6.7%) received the maximum score of 20 from both raters, and 6 students (20.0%) had a mean posttest score of 19 or higher; the highest mean posttest score in M32 was 16.5. Taken separately, Rater 1 gave 20 to five M29 posttests and Rater 2 to two. Scores at or near the ceiling leave little room for further gain, which would tend to understate rather than exaggerate the gains of the highest-scoring M29 students, but the scale cannot distinguish among the strongest writers in that class. The rank-based and pretest-adjusted analyses in Section 4.7 are less sensitive to this than a comparison of raw means alone.

Figure 6 shows the full frequency distribution of gains, complementing Table 5 and the individual score plots.

Figure 6. Distribution of individual gains



Note. Bars show student counts at each recorded gain. Gain = posttest − pretest on the two-rater mean; n = 30 per class.

Robustness of the Comparison

Table 6 repeats the comparison of gains on each rater's scores separately and adjusts for the pretest. The direction and general size of the difference did not depend on which rater's scores were used, but the size did vary: the gain difference was 3.00 points on Rater 1's scores and 2.27 points on Rater 2's, a gap of 0.73 points that arises largely from Rater 2's higher marks for the M32 posttests (Section 4.1). Rater 1 was the class instructor, so this is a reason to treat the two-rater mean as the primary estimate and the range as an indication of the uncertainty that scoring adds. After adjustment for the pretest, the portfolio class scored 2.68 points higher at posttest, 95% CI [2.01, 3.35], $t(57) = 7.96$, $p < .001$; the pooled pretest slope was 0.94, and there was no clear evidence that slopes differed between the classes ($p = .10$).

Table 6. Comparison of gains by rater and with adjustment for the pretest

Scores analysed	M29 gain	M32 gain	Difference in gain [95% CI]	Hedges' g	Pretest-adjusted difference [95% CI]
Mean of two raters (primary)	4.53	1.90	2.63 [1.97, 3.30]	2.03	2.68 [2.01, 3.35]
Rater 1 only	4.53	1.53	3.00 [2.25, 3.75]	2.05	3.02 [2.26, 3.78]
Rater 2 only	4.53	2.27	2.27 [1.55, 2.99]	1.61	2.42 [1.70, 3.13]

Note. Difference = M29 minus M32. The pretest-adjusted difference is the class difference in posttest scores from a regression of posttest on class and pretest. Intervals are conventional student-level 95% confidence intervals, unadjusted for class clustering.

DISCUSSION

Principal Findings

The central result is a larger improvement in the portfolio class than in the traditional class. M29 gained 4.53 points on average and M32 gained 1.90, a difference of 2.63 points on a 20-point scale. The result appears in the class means, in the distributions of individual gains, and in the paired score display, and it persisted when each rater's scores were analysed separately (2.27 to 3.00 points) and after adjustment for the pretest (2.68 points). Every student in both classes improved, and 60% of the portfolio class gained four points or more, against none of the traditional class. The raters agreed closely (ICC(2,1) = .96), so the result is unlikely to reflect idiosyncratic scoring by either rater, although the size of the difference was somewhat larger on the instructor's scores than on the second rater's.

The findings are consistent with the instructional rationale for portfolios. Both classes received teaching and corrective feedback, and the class that also kept a cumulative record of assignments and feedback showed greater growth in overall paragraph writing. They do not show that the portfolio arrangement caused the difference. With one class per condition, the portfolio condition is confounded with everything else that may have distinguished M29 from M32, such as the composition and motivation of the students, the timetable of each class, and the dynamics of each group with its instructor, and these cannot be separated from the arrangement with the present data (Section 7).

Continuity Across Assignments

One interpretation of the result is that retaining earlier work made corrective feedback easier to revisit when later tasks raised similar problems. A correction that remains available can be consulted after the immediate grading event has passed, and a portfolio can help link separate assignments into a longer learning sequence. This interpretation fits the instructional contrast between the classes and the conceptual framework in Section 2, but it is an interpretation and not a finding of this study: no record was kept of whether or how students consulted their portfolios, and the study did not code revisions or compare the errors in successive drafts. Other explanations, such as differences in how engaged the two classes were, are equally compatible with the scores.

The range of gains in M29, from 1.5 to 8.0 points, shows that students in the portfolio class benefited to different degrees. Why this was so cannot be established from total scores. An analysis of drafts that distinguishes copying a correction from applying a principle to a new context, together with students' own accounts of how they used the portfolio, would be a useful way of studying it.

The Size of the Improvement

The difference of 2.63 points in mean gain is meaningful in practical terms. It equals 13% of the full scale, and it appeared alongside gains of four points or more for 60% of the portfolio class and for none of the traditional class. The standardized difference, Hedges' g = 2.03, is very large by conventional benchmarks and should be read with caution. It reflects the separation between the class means relative to the spread of individual gains, and gains within each class were fairly uniform, particularly in M32, where the standard deviation of gains was 0.71. Standardized effects depend on the measure and on how the denominator is defined, and they are likely to be inflated in small, non-randomized comparisons of two classes, so the raw difference on the rubric scale and the individual-level distributions provide the most transparent account of the result. Comparisons with standardized effects from other studies are most informative when the designs and denominators are aligned (Lakens, 2013).

Progress under Traditional Instruction

The traditional class also improved, and this is an encouraging result in its own right. All 30 students in M32 gained points, and the class mean rose by 1.90. Part of this rise may reflect the difference between a diagnostic pretest and a midterm posttest, and part the effect of eight weeks of instruction; the design cannot apportion it. Regular instruction with corrective feedback and redrafting is associated with growth in paragraph writing, and the portfolio class improved by more than that.

Overall Writing Quality and the Ceiling of the Scale

The rubric combines five aspects of writing into one total, and higher totals indicate better overall paragraph writing under this scoring arrangement. A student can gain points through clearer organization, stronger support for ideas, more accurate grammar, richer vocabulary, or more consistent mechanics, and the rubric's equal weighting means that ideas and organization account for two fifths of the total and language for three fifths. The raters recorded totals only, so the study cannot show in which dimension the gains were concentrated; recording dimension scores would answer this and is a priority for future work.

Two portfolio students received the maximum score of 20 from both raters, and 6 had a mean posttest score of 19 or higher. This shows that the strongest writers in the class reached the top of the scale, which limits the measurement of progress at that end; a more demanding task or scoring scheme would extend it. Because a ceiling compresses the gains of the highest scorers, it would tend to understate, not exaggerate, the portfolio class's advantage.

Relationship to Earlier Scholarship

The findings fit the broader instructional position that feedback should inform subsequent action and not end with the return of a marked paper. This position links the present classroom comparison with formative assessment and feedback literacy scholarship (Black & Wiliam, 1998; Carless & Boud, 2018; Nicol & Macfarlane-Dick, 2006). A portfolio is one practical way of acting on it, since it keeps earlier texts and responses at hand when a new task begins. The study also complements electronic portfolio research (Pourdana & Tavassoli, 2022) and research on writing conceptions and approaches (Gebrekidan & Zeru, 2023) by adding a small set of scores from overall paragraph writing in a university EFL class. Some of the studies reviewed in Section 2 involved several classes (Abduljawad, 2024) or random assignment of learners (Fathi et al., 2020); the present study should be read as a modest, class-level contribution alongside them.

Pedagogical Implications

Designing a Manageable Portfolio Sequence

The practical lesson of the study is to make the connection between assignments explicit. A teacher introducing portfolios can ask students to keep an initial text, the corresponding feedback, and a revised version together. Before beginning a later task, students can identify a small number of recurring issues to watch for in their next paragraph. The purpose is for the portfolio to serve learning during the course and not to become a collection assembled for inspection at the end.

The sequence works best when it stays manageable. Extensive reflection after every minor correction can create work without a proportionate gain in learning, whereas a focused response that identifies the problem, explains a change, and checks whether the problem recurs is quicker and more useful. The appropriate level of detail depends on students' language resources and the course workload, so instructors can begin with clearly explained expectations and adjust as they see how students respond.

Balancing Feedback on Language and Meaning

A portfolio approach should attend to both linguistic accuracy and the development of ideas. If students understand revision as correcting spelling and verb forms only, improvement in coherence and support may receive less attention. Broad comments about meaning, on the other hand, can be hard to act on without concrete guidance. Teachers can connect comments to the relevant rubric dimension and illustrate what a successful response would accomplish, so that students understand the reason for a change and can prioritize among several possible corrections.

The five equally weighted dimensions of the rubric provide a helpful organizing structure. They do not require equal comment volume on every assignment: a student's needs and the purpose of a task may justify concentrating feedback on selected issues. A portfolio makes the process visible by linking feedback with later work, and a periodic check by the instructor confirms that students are using it.

Using Exemplars and Brief Reflection Notes

A useful extension is to compare a short anonymous example with the same assessment criteria applied to students' own work. An instructor might present two versions of a supporting sentence and ask which explains the topic more clearly, and students can then look for a comparable issue in their own paragraph. This links evaluation with revision and gives learners a reason for making a change, while leaving the content of the paragraph in their hands.

The portfolio can keep a record of such decisions through a brief note attached to each revision, in which a student identifies the original weakness, describes the intended improvement, and points to the changed passage. The notes show the instructor how students interpreted a comment and where additional explanation would help. When a student repeatedly changes a verb form in an unsuccessful way, for example, the note and the drafts together help show whether the difficulty lies in grammatical knowledge, in interpreting the correction, or in the demands of the new sentence. A final corrected draft alone would not reveal this.

Assessment Conditions and Workload

Portfolio-supported practice and outcome assessment serve different purposes. During learning, consultation of earlier work, revision, and feedback are encouraged, and during an assessment designed to measure independent performance, the access conditions can be stated and applied consistently. Keeping the two purposes clear lets instructors support development while still learning whether students can apply what they have learned to a new paragraph.

A portfolio does not remove the time needed to respond thoughtfully to student writing, but it can help organize that time around recurring needs and not repeated isolated corrections. Sampling selected revisions,

checking targeted issues, and using concise criterion-based comments are practical options that instructors can adapt to their own classes.

LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH

The study was carried out in two intact classes in one university setting, with one class in each instructional condition and one instructor. The results therefore describe what happened in these classes: the portfolio class improved more than the traditional class, and the improvement was broad. Because the condition was applied to whole classes, class-level differences in student composition, motivation, timetable, and classroom dynamics are confounded with it, and the observed difference cannot be attributed exclusively to the portfolio arrangement. The tests and confidence intervals treat students as independent observations; with one class per condition the variation between classes cannot be estimated, so these statistics are descriptive of the sample. The classes were not equivalent at the start, although the pretest difference was not statistically significant, and the instructor knew the purpose of each class's routine. A design with several classes per condition, ideally with random assignment of classes and instructors, would estimate class-to-class variation and extend the findings beyond this setting. Conclusions in this paper should remain associational and class-specific unless such data become available.

Measurement has its own limits. Scores are the mean of two human raters, and the raters agreed closely, but they did not agree perfectly: the size of the difference between the classes varied by about 0.7 points between the two raters' scores, and Rater 1 was the instructor of both classes and may have recognised the class of a script. Totals were recorded without dimension scores, so the study cannot say whether gains were concentrated in content, organization, grammar, vocabulary, or mechanics. The pretest was diagnostic and the posttest a midterm examination, with prompts that differed in topic, genre, scaffolding, and word limit and were not equated for difficulty, so the absolute gains in both classes are uncertain even though the comparison between classes used the same tests. A ceiling was approached by some portfolio students at posttest. No delayed posttest was given, so retention and transfer are untested.

Finally, the proposed mechanism, that keeping a portfolio improved review and use of feedback, was not measured. No portfolio-use records, revision analyses, or student accounts were collected, and statements in this paper about how portfolios may work are interpretations and not results. The data are secondary use of routine course assessments, which raises the ethical considerations noted in the Declarations.

The study points to several directions that build on its findings:

1. **Replication across classes.** Running the comparison in several classes per condition, with instructors and timetables varying, or adding archived cohorts of the same course, would test how far the result extends and would clarify the contribution of the portfolio arrangement itself.
2. **Feedback use.** Coding linked drafts, comments, and later revisions for a sample of students (for example as successful revision, unsuccessful revision, no response, or a change unrelated to the comment) would show how students use the portfolio and whether later texts repeat fewer of the errors addressed earlier.
3. **Student perspectives.** Brief interviews or reflective accounts would show what students find useful, confusing, or burdensome about portfolios, and could be analysed alongside the score data.
4. **Task, rubric, and scoring design.** Comparable writing prompts counterbalanced across occasions, anonymised scripts scored by raters unaware of class and occasion, dimension scores retained for every script, a delayed posttest, and tasks with more room at the top of the scale would allow progress to be measured more securely, for the strongest writers as well as for the rest of the class.

These steps would keep the practical focus of the present study and add evidence about how and for whom portfolio-based instruction works best.

CONCLUSION

This study compared paragraph writing development in two university EFL classes that both received instruction and corrective feedback. The class that kept its assignments and feedback in a cumulative portfolio improved from 11.93 to 16.47 on a 20-point rubric, a mean gain of 4.53 points, while the traditional class improved from 11.12 to 13.02, a mean gain of 1.90 points. The 2.63-point difference in gains was reflected across students: every student in both classes improved, and 60% of portfolio students gained four points or more, against none in the traditional class. The difference was present for each of two raters whose scores agreed closely and after adjustment for the pretest.

The educational message concerns the value of connecting written assignments through retained feedback and subsequent review. In this study, keeping a cumulative portfolio was associated with greater growth in overall paragraph writing than traditional instruction with corrective feedback. Because the comparison involves one class per condition, the association is specific to these two classes and does not establish that the portfolio caused the difference. The approach is practical, requires no special technology, and can be tried in existing university writing courses, particularly when expectations for using the portfolio are explained clearly and the feedback is tied to the rubric dimensions students are working on.

The findings invite further work with additional classes, dimension-level scores, evidence of how students use their feedback, and delayed assessments. They already provide a clear rationale for making feedback use a visible and systematic part of university EFL writing instruction.

Declarations

Ethics approval and consent. The pretest and posttest were routine course assessments (a diagnostic test at the start of the course and the midterm examination), taken in classes taught by the first author, a faculty member at Taibah University. Students were not told when they took the tests that the scores might later be used for research. The data analysed are anonymised scores, and student names were removed from the analytic file. After the course, the students were sent an online form about the use of their anonymised scores for research, and all agreed. The institution permitted the use of the data on condition that no names were used, and none are reported.

Funding. The authors received no funding for this study.

Competing interests. The authors declare no competing interests.

Data availability. The anonymised pretest and posttest marks of both raters for all 60 students are reported in Appendix A, which is sufficient to reproduce every analysis in this paper. The analysis code is available from the corresponding author on request.

Use of AI tools. No generative AI tool was used to score the writing samples or to analyse the data.

REFERENCES

1. Abdel Latif, M. M. M., Alsuhaibani, Z., & Alsahil, A. (2024). Matches and mismatches between Saudi university students' English writing feedback preferences and teachers' practices. Assessing Writing, 61, Article 100863. https://doi.org/10.1016/j.asw.2024.100863
2. Abduljawad, S. (2024). Teachers to learners: Portfolio, please! New techniques of portfolio assessment in ESL classrooms. International Journal of Learning, Teaching and Educational Research, 23(4), 34–51. https://doi.org/10.26803/ijlter.23.4.3
3. Alam, M. J., & Aktar, T. (2019). Assessment challenges & impact of formative portfolio assessment (FPA) on EFL learners' writing performance: A case study on the preparatory English language course. English Language Teaching, 12(7), 161–172. https://doi.org/10.5539/elt.v12n7p161
4. Alshahrani, A., & Storch, N. (2014). Investigating teachers' written corrective feedback practices in a Saudi EFL context: How do they align with their beliefs, institutional guidelines, and students'

preferences? Australian Review of Applied Linguistics, 37(2), 101–122. https://doi.org/10.1075/aral.37.2.02als

5. Alshahrani, A., & Storch, N. (2025). Investigating the effectiveness of scaffolded feedback on EFL Saudi students' writing accuracy: A longitudinal classroom-based study. Assessing Writing, 63, Article 100910. https://doi.org/10.1016/j.asw.2024.100910
6. 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
7. Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet, 327(8476), 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8
8. Boumediene, H., Berrahal, F. K., & Harji, M. B. (2016). The effectiveness of portfolio assessment on EFL students' writing performance: The case of third year secondary students in Algeria. Academic Journal of Interdisciplinary Studies, 5(3 S1), 119. https://doi.org/10.5901/ajis.2016.v5n3s1p119
9. Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. Assessment & Evaluation in Higher Education, 43(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354
10. Farahian, M., & Avarzamani, F. (2018). The impact of portfolio on EFL learners' metacognition and writing performance. Cogent Education, 5(1), Article 1450918. https://doi.org/10.1080/2331186X.2018.1450918
11. Fathi, J., Derakhshan, A., & Safdari, M. (2020). The impact of portfolio-based writing instruction on writing performance and anxiety of EFL students. Polish Psychological Bulletin, 51(3). https://doi.org/10.24425/ppb.2020.134729
12. Ferris, D. (1999). The case for grammar correction in L2 writing classes: A response to Truscott (1996). Journal of Second Language Writing, 8(1), 1–11. https://doi.org/10.1016/S1060-3743(99)80110-6
13. Ferris, D. (2010). Second language writing research and written corrective feedback in SLA: Intersections and practical applications. Studies in Second Language Acquisition, 32(2), 181–201. https://doi.org/10.1017/S0272263109990490
14. Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement, 33(3), 613–619. https://doi.org/10.1177/001316447303300309
15. Gebrekidan, H., & Zeru, A. (2023). Effects of portfolio-based assessment on EFL students' conceptions and approaches to writing. Cogent Education, 10(1), Article 2195749. https://doi.org/10.1080/2331186X.2023.2195749
16. Hamp-Lyons, L., & Condon, W. (2000). Assessing the portfolio: Principles for practice, theory, and research. Hampton Press.
17. Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. Language Teaching, 39(2), 83–101. https://doi.org/10.1017/S0261444806003399
18. Kang, E., & Han, Z. (2015). The efficacy of written corrective feedback in improving L2 written accuracy: A meta-analysis. The Modern Language Journal, 99(1), 1–18. https://doi.org/10.1111/j.1540-4781.2015.12187.x
19. Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. Frontiers in Psychology, 4, Article 863. https://doi.org/10.3389/fpsyg.2013.00863
20. Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012
21. Lam, R. (2017). Taking stock of portfolio assessment scholarship: From research to practice. Assessing Writing, 31, 84–97. https://doi.org/10.1016/j.asw.2016.08.003
22. Liu, K., & Storch, N. (2023). Second language learners' engagement with written feedback. Australian Review of Applied Linguistics, 46(1), 4–28. https://doi.org/10.1075/aral.20029.liu
23. 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

24. Obeiah, S. F., & Bataineh, R. F. (2016). The effect of portfolio-based assessment on Jordanian EFL learners' writing performance. Bellaterra Journal of Teaching & Learning Language & Literature, 9(1). https://doi.org/10.5565/rev/jtl3.629
25. Oshima, A., & Hogue, A. (2006). Writing academic English (4th ed.). Pearson/Longman.
26. Pourdana, N., & Tavassoli, K. (2022). Differential impacts of e-portfolio assessment on language learners' engagement modes and genre-based writing improvement. Language Testing in Asia, 12, Article 7. https://doi.org/10.1186/s40468-022-00156-7
27. Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420
28. Truscott, J. (1996). The case against grammar correction in L2 writing classes. Language Learning, 46(2), 327–369. https://doi.org/10.1111/j.1467-1770.1996.tb01238.x
29. Wang, L., & He, C. (2020). Review of research on portfolios in ESL/EFL context. English Language Teaching, 13(12), 76–82. https://doi.org/10.5539/elt.v13n12p76

APPENDIX A. ANONYMISED SCORES OF BOTH RATERS

Scores out of 20 for each student at pretest and posttest from Rater 1 (R1) and Rater 2 (R2). Students are identified by a code assigned in class-list order; names are not retained. The score analysed for each student is the mean of R1 and R2.

M29 ID	Pre R1	Pre R2	Post R1	Post R2	M32 ID	Pre R1	Pre R2	Post R1	Post R2
M29-01	15	16	20	18	M32-01	14	13	17	16
M29-02	14	13	20	19	M32-02	13	14	16	15
M29-03	8	8	14	14	M32-03	7	8	8	9
M29-04	9	10	11	12	M32-04	8	9	9	10
M29-05	13	14	17	16	M32-05	12	11	14	13
M29-06	8	7	16	15	M32-06	6	7	8	9
M29-07	7	8	10	11	M32-07	11	12	12	12
M29-08	11	12	16	16	M32-08	12	13	13	14
M29-09	11	10	14	15	M32-09	10	9	11	12
M29-10	10	11	15	16	M32-10	9	10	11	12
M29-11	13	12	18	17	M32-11	14	13	16	16
M29-12	12	13	17	16	M32-12	11	12	12	13
M29-13	9	9	13	14	M32-13	12	11	13	14
M29-14	12	11	19	18	M32-14	7	8	10	11
M29-15	13	14	20	20	M32-15	12	11	14	14
M29-16	15	14	20	19	M32-16	14	13	16	16
M29-17	10	11	17	18	M32-17	9	10	11	12
M29-18	11	12	19	18	M32-18	9	10	11	12
M29-19	14	13	16	17	M32-19	13	12	15	14
M29-20	13	12	17	18	M32-20	12	11	14	15
M29-21	12	11	14	15	M32-21	11	11	12	13
M29-22	11	12	14	15	M32-22	13	12	15	15
M29-23	16	16	18	17	M32-23	13	12	14	15
M29-24	14	13	16	17	M32-24	12	11	11	13
M29-25	8	7	11	11	M32-25	8	7	8	9
M29-26	15	15	19	18	M32-26	14	13	15	15
M29-27	15	15	18	18	M32-27	14	13	17	16
M29-28	14	13	20	20	M32-28	14	13	15	16
M29-29	13	14	19	19	M32-29	13	12	14	15
M29-30	12	12	16	17	M32-30	10	9	11	12

Note. R1 = Rater 1 (first author, class instructor); R2 = Rater 2 (second author).

Appendix B. Writing prompts and rubric

Pretest (Writing 103, 45 minutes). One paragraph of 80–100 words entitled “My University Major”. Students could use these questions: What is your major? Why did you choose this major? What do you like about it? What is difficult about it? What job would you like to have in the future? Instructions: write one paragraph of about 80–100 words; try to include a topic sentence, two or three supporting ideas, examples or details, and a concluding sentence; do not worry if your English is not perfect; write as much as you can.

Posttest (Writing 103, 45 minutes). Paragraph writing: write one paragraph of 100–120 words. Include a clear topic sentence, supporting details, and a concluding sentence. Use suitable transition signals. Topic: The effects of eating too much fast food. Explain two or three effects on a person’s health and daily life.