Journal of Creative Arts, Vol. 2, No. 2 (Special Theme 1), 2025
e-ISSN: 3030-6086
Page: 125-143

AI Integration in Writing Skills of MUET Students: A Quasi-Experimental Study

Peter Ong
Open University Malaysia (OUM)
Corresponding Author : peterong222@gmail.com

Received: November 10th, 2025
Accepted: December 19th. 2025
Published: December 22nd, 2025

ABSTRACT

This quasi-experimental study examines the effects of artificial intelligence (AI)-assisted feedback on Malaysian University English Test (MUET) writing performance among Form 6 students. A sample of 80 students from two Malaysian secondary schools who were matched in their Lower 6 examination results was randomly assigned to experimental (n=40, AI-assisted feedback) and control groups (n=40, traditional teacher feedback). Both groups received the same classroom writing instruction in a four-week intervention, and only the feedback provision channels are different. The experimental group employed AI writing supports (Grammarly, Hemingway Editor, QuillBot) which generated real-time automatic feedback on grammar, cohesio, organization and lexis diversity. Conventional written teacher feedback was provided to the control group 48-72 hours after attending standard MUET preparation programme. Data were analyzed using a pretest-posttest design with independent samples t-test. The experimental group exhibited statistically significant differences on posttest writing scores (Mexperimental=91.33, SD=4.39; Mcontrol=82.78, SD=3.01), t(78)=10.16, p<.. 001, two-tailed with a very large effect size (Cohen's d=2.27). At component level, all MUET Writing criteria (Task Fulfillment, Organization, Language Use and Mechanics) had significant improvement but the effects were highest on Organization (d=1.24) and Mechanics (d=1.26). Results show that AI-enabled feedback is beneficial for English as a Second Language (ESL) writing in high-stakes assessment situations, as long as it is provided strategically. The research provides empirical data to back up the integration of AI for MY English language. Further longitudinal retention studies are needed to determine the roles of student perceptions and attitudes towards AI tools as well as optimal blended learning designs that combine those feedbacks with others from humans, which can be implemented in various educational settings or populations.

Keywords: Artificial Intelligence, writing skills, MUET, ESL education, quasi-experimental design

PENERBIT PRESS
UNIVERSITI TEKNOLOGI MARA

1. INTRODUCTION

1.1 Background and Context

The introduction of Artificial Intelligence (AI) into educational practice brought a revolutionary change to the language teaching and learning approaches (Chen et al., 2024; Fitria, 2021). Such AI-supported writing tools as Grammarly, ChatGPT and automated essay evaluation systems are able to give students instant, specific feedback in such aspects as grammar accuracy, textual coherence and lexical diversity (Pratama & Sulistiyo, 2024; Wang et al., 2023). Writing is a significant aspect of the assessment and it contributes to the Malaysian University English Test (MUET) which is an important yardstick for measuring students' English language proficiency in Malaysia. As MUET has high stakes consequences for university admission, novel writing pedagogies need rigorous empirical scrutiny (Abdul Razak & Md Yunus, 2021).

1.2 Research Problem

Typically-taught writing practices tend to result in long delays between student composition and teacher response, meaning that students have few chances for immediacy of error correction and the revision process (Hyland & Hyland, 2019; Lee, 2020). Furthermore, teacher feedback is often inconsistent between assessors due to differences in interpreting rubrics and teaching philosophies (Lam, 2022; Zhang & Hyland, 2023). There drawbacks are especially pronounced in low-resource educational environments where teach-student ratios make it impossible to provide personalized attention (Ferris, 2021). The current study seeks to determine the possibility of AI-generated feedback as a complementary or alternative means of producing traditional feedback mechanisms in preparing for MUET writing.

1.3 Research Objectives

This study aims to:

1. Evaluate the effectiveness of AI-assisted feedback compared to traditional teacher feedback on MUET writing performance
2. Examine the magnitude of improvement differences between intervention and control conditions
3. Provide empirical evidence for AI integration in Malaysian ESL writing instruction

1.4 Research Questions and Hypotheses

Research Question: Does AI-assisted feedback produce significantly different MUET writing scores compared to traditional teacher feedback?

Null Hypothesis (H_0): There will be no statistically significant difference in post-intervention writing scores between students receiving AI-assisted feedback and those receiving traditional teacher feedback.

PENERBIT PRESS
UNIVERSITI TEKNOLOGI MARA

Alternative Hypothesis (H_1): Students receiving AI-assisted feedback will demonstrate significantly different writing scores compared to those receiving traditional teacher feedback.

1.5 Significance of the Study

This study adds to the emerging body of research on AI efficacy in the ESL setting (Escalante et al., 2023; Shang, 2024). This study used a controlled quasi-experimental design to offer strong evidence of AI's pedagogical merit in MUET preparation. The results have practical implications for teachers who wish to improve the manner in which feedback is given; regarding policy making for the adoption of educational technology at institutional level; and, policymakers from whom it can inform when developing an English curriculum for education in Malaysia (Taskiran et al., 2024).

1.6 Scope and Delimitations

The current study involved only 80 Form 6 students at two Malaysian secondary schools and focuses on those who are preparing for the MUET writing tests. The intervention lasted 4 weeks, and was specifically focused on writing skills (language production), so that there were no speaking, listening or reading invitations. Subjects were naive to AI-based writing instruction, allowing for novelty of treatment. The experiment only investigates new immediate post-intervention as opposed to long-term or transfer effects.

1.7 Limitations

Several constraints limit generalizability. First, the two-school sample limits external validity to like institutional settings. Second is that the short intervention period might not be suitable for representing long-term l earning trajectories. Third, although AI tools are good at localized correction (grammar, mechanics), they posses only limited ability to assess higher order concerns such as quality of argumentation encountered, depth of critical thinking displayed, and sophistication of rhetoric used in the written piece (Makwana, 2025; Quratulain et al., 2025). Fourth, the research has not considered heterogeneity in digital literacy and technology acceptance as potential moderators of AI effectiveness.

2. LITERATURE REVIEW

2.1 AI: Applications for Language Education

AI applications have spread throughout the learning landscape delivering personalised learning and scaling feedback delivery (Holmes et al., 2019; Luckin et al., 2016). In the field of language learning, AI has many potential applications such as intelligent tutoring systems or automated writing evaluation (AWE) to analyze linguistic features of an individual's text and suggest real-time feedback for improvement (Warschauer & Grimes, 2008). Recent syntheses show that AI-enhanced language teaching (AILT), more specifically AI-mediated language learning (technologies), holds potential for writers who wish to develop fluency, accuracy, and sophistication in vocabulary (Pratama & Sulistiyo, 2024; Wang et al.

PENERBIT PRESS
UNIVERSITI TEKNOLOGI MARA

2.2 Automated Writing Evaluation and Feedback (AWE)

Researchers have traditionally been interested in the effects of AWE upon writing through feedback dieses (e.g., Warschauer & Healey, 1998; Aydelott et al., 2005). Automated writing evaluators use natural language processing algorithms to score aspects of the quality of writing, such as mechanical accuracy of language, syntactic complexity, coherence cues and lexical diversity (Shermis & Burstein, 2013; Wilson & Roscoe, 2020). There is mixed evidence for AWE effectiveness: improvements in surface-level accuracy (spelling, grammar, and punctuation) have been well-documented but proof of improved quality at the level of rhetorical discourse construction remains elusive (Stevenson & Phakiti 2019). Notably, Escalante et al. (2023) found that the groups to which AI-derived feedback was provided saw considerable enhancements in essay organization and ability for self-revision among university students more than community of practice.

2.3 ESL Writing Problems 2.3.1 Writing in ESL Settings

Writing is especially problematic for ESL writers with a narrow language repertoire and L1 transfer issues (Jacobsen, 2014) who are still struggling to be exposed to L2 discourse traditions and expectations (Hyland, 2019). In MUET writing tasks, which are tests used in the Malaysian context, candidates need to write coherent and well-structured essays with evidence of accurate grammar, appropriate register and persuasive argument (Malaysian Examinations Council, 2021). While useful for attending to higher-order issues, the feedback provided by traditional teacher comments is frequently untimely (i.e., delayed) in terms of being able to contribute usefully to immediate revision work and can vary widely in focus and specificity (Ferris, 2021; Lee, 2020).

2.4 Computer-aided feedback in ESL Writing

Theoretical evidence for L2 written AI feedback Despite the limitations in established motivational and cognitive theories, there are theoretical grounds for considering the use of AI feedback in L2 academic writing (cf. Quratulain et al. (2025) explored AI writing assistants and their impact on Pakistani undergraduates where compared to the control groups, experimental units were more accurate in touching upon the organisation level is as well as self-revised. But participants raised the issue of dependence and academic honesty. Likewise, Makwana (2025) investigated AI-generated and teacher-written feedback as factors within formative L2 writing assessments from Indian ESL student writers, finding that when looking for scored response types, the two types of feedback were similar in terms of effectiveness on some measures of writing scores but also problematic with regard to contextual appropriateness and ethical transparency.

2.5 The Type of MUET and English Language Assessment in Malaysia

MUET is the major gatekeeping in university entry for Malaysian, which tests their listening, speaking, reading and writing at six band levels (Abdul Razak & Md

PENERBIT PRESS
UNIVERSITI TEKNOLOGI MARA

Yunus, 2021). There is a crucial written element in which candidates are required to produce a piece of extended discourse displaying linguistic capacity and rhetorical skill.

2.6 Quasi-Experimental Studies of Educational Technology

Quasi-experimental designs facilitate causal inference to be made in the naturalistic learning environment when random assignment is not feasible, or ethical (Creswell, 2021). These designs use available groups with appropriate statistical controls to mimic experimental conditions. In educational technology research, quasi-experimental designs trade internal with ecological validity; they allow observation of the effect in real classroom settings.

2.7 Gaps in Existing Research

A number of researches have studied the effectiveness of AI for grammar correction which is well documented, but in a strictly controlled assessment context like MUET, empirical studies on the effects on student writing quality are sparse. Previous studies frequently use a convenience sample, do not have adequate control groups, or rely on self-reported perceptions instead of objective performance. This study fills these gaps through the controlled contrast of AI-assisted and traditional feedback on a set of validated MUET writing assessments in test-like conditions.

2.8 Theoretical Framework

This article combines two theoretical perspectives: the Automated Writing Evaluation Framework (Warschauer & Grimes, 2008) and Feedback Intervention Theory (Kluger & DeNisi, 1996).

Automated Writing Evaluation Framework theorises feedback from AI as a technological scaffolding that enhances human teaching with instantaneous, uniform and in-depth linguistic analysis (Warschauer & Grimes, 2008). AWE systems do operationalize the notion of writing quality, by using computer-linguistics to find surface errors and flag incoherence as well as style. This is a framework that promotes AI as something additive, not substitutive to writing pedagogy.

According to the Feedback Intervention Theory, an effective feedback is one that focuses learners attention on task-level processes rather than self evalucation concerns (Kluger & DeNisi, 1996). Powerful feedback is specific, immediate, and actionable, allowing learners to discern discrepancies between where they are vs. where they want to be (or need to be). Immediacy and specificity in AI feedback may have the potential to improve upon these conditions, offering granular guidance at the moment of reading.

PENERBIT PRESS
UNIVERSITI TEKNOLOGI MARA

Journal of Creative Arts, Vol. 2, No. 2 (Special Theme 1), 2025
e-ISSN: 3030-6086
Page: 125-143

Conceptual Model:

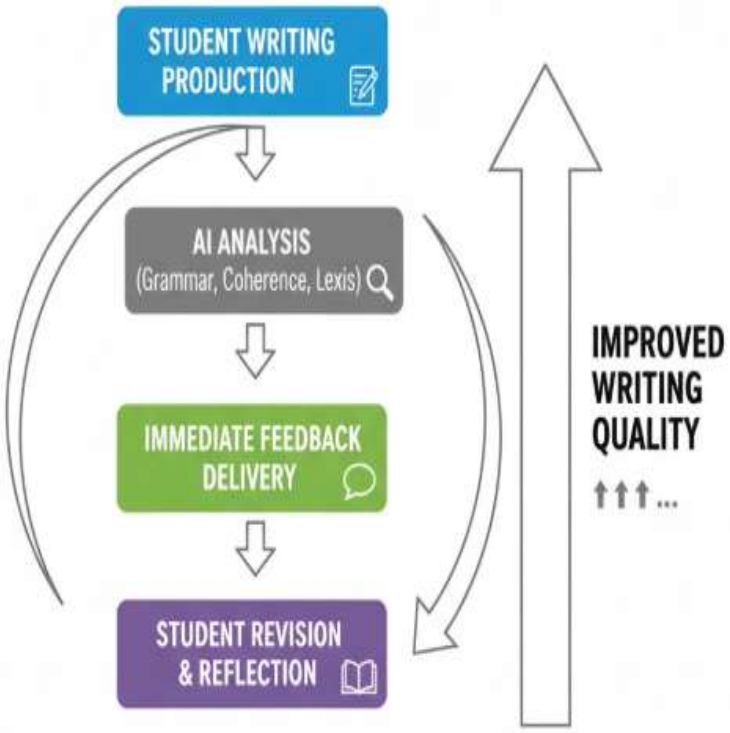


Figure 1: Conceptual model of AI-assisted feedback mechanism

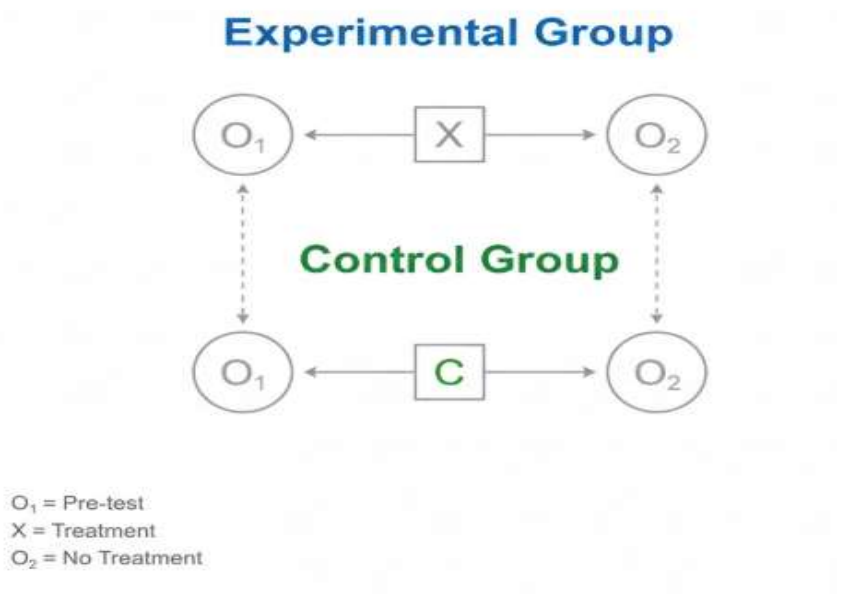
This dual-framework approach recognizes that AI tools function both as technological artifacts with specific affordances (AWE perspective) and as pedagogical interventions that must align with cognitive principles of learning (FIT perspective). The integration of these frameworks guided intervention design, ensuring AI tools provided linguistically sophisticated and pedagogically appropriate feedback.

3. METHODOLOGY

3.1 Research Design

PENERBIT PRESS
UNIVERSITI TEKNOLOGI MARA

This study employed a quasi-experimental pretest-posttest control group design to investigate AI feedback effectiveness on MUET writing performance. The design structure is represented as:



Where O_1 = pretest, O_2 = posttest, X = AI-assisted feedback intervention, C = traditional teacher feedback. This design enables causal inference while maintaining ecological validity within authentic classroom settings.

3.2 Participants and Sampling

Participants comprised 80 Form 6 students purposively selected from two Malaysian secondary schools based on comparable Lower 6 final examination performance (M=68.5%, SD=4.2%). Purposive sampling ensured baseline equivalence across critical demographic and academic variables. Following school-level selection, students were randomly assigned to experimental (n=40) or control (n=40) conditions using computer-generated random numbers.

Inclusion Criteria:

- Currently enrolled in Form 6 (pre-university level)
- English as Second Language learner
- No prior exposure to AI-integrated writing instruction
- Comparable academic performance (±5% variance in Lower 6 results)

Demographic Characteristics:

- Age range: 17-18 years (M=17.4, SD=0.51)
- Gender distribution: 48 female (60%), 32 male (40%)

PENERBIT PRESS
UNIVERSITI TEKNOLOGI MARA

- L1 backgrounds: Malay (65%), Chinese (22%), Tamil (13%)

3.3 Instruments

MUET Writing Assessment Tool: The study employed standardized MUET Task 2 writing prompts (argumentative essays, 350+ words) developed by the Malaysian Examinations Council. Essays were scored using official MUET analytical rubrics across four criteria: Task Fulfillment, Organization, Language, and Mechanics (each weighted 25%, total 100 points). Two trained raters independently scored all essays; inter-rater reliability coefficients exceeded .85 (Cohen's kappa), indicating substantial agreement.

AI Feedback Tools (Experimental Group):

- **Grammarly Premium:** Grammar, punctuation, spelling correction; style and tone suggestions
- **Hemingway Editor:** Sentence complexity analysis; readability scoring
- **QuillBot:** Paraphrasing suggestions; vocabulary enhancement

Traditional Feedback (Control Group): Teachers provided handwritten margin comments and end-of-essay summative feedback addressing content, organization, language use, and mechanics, consistent with typical MUET preparation practices.

3.4 Procedures

Week 0 (Baseline): All participants completed identical pretest writing tasks under examination conditions (90 minutes, no feedback provision). Essays were scored by independent raters blind to group assignment.

Weeks 1-4 (Intervention): Both groups received equivalent writing instruction (3 sessions/week, 60 minutes/session) covering MUET essay structure, argumentation strategies, and language use. Groups differed exclusively in feedback delivery:

Experimental Group: Students drafted essays individually, submitted digital copies via learning management system, received AI-generated feedback within 24 hours, and revised based on AI suggestions. Teachers monitored progress but did not provide direct feedback.
Control Group: Students drafted essays, submitted hard copies to teachers, received handwritten feedback within 48-72 hours, and revised based on teacher comments.
Both groups completed identical writing assignments (4 major essays, 2 timed practices).

Week 5 (Post-intervention): All participants completed parallel posttest writing tasks under examination conditions. Posttests employed alternate MUET prompts of equivalent difficulty (as determined by expert panel review) to minimize practice effects.

PENERBIT PRESS
UNIVERSITI TEKNOLOGI MARA

3.5 Statistical Analysis

Data were analyzed using IBM SPSS Statistics 28.0. Prior to hypothesis testing, assumptions were verified:

1. **Normality:** Shapiro-Wilk tests indicated normal distributions for both pretest (experimental: W=.97, p=.31; control: W=.98, p=.52) and posttest (experimental: W=.96, p=.18; control: W=.97, p=.35) scores.
2. **Homogeneity of Variance:** Levene's test assessed variance equality between groups. For pretest, F(1,78)=0.52, p=.47; for posttest, F(1,78)=3.64, p=.06, both non-significant at α=.05, supporting equal variance assumption.

Primary Analysis: Independent samples t-tests compared mean posttest scores between experimental and control groups. Alpha level was set at .05 (two-tailed). Effect sizes were calculated using Cohen's d to quantify practical significance (Cohen, 1988).

Justification: Independent samples t-test represents the appropriate parametric test for comparing means between two independent groups when assumptions of normality and homogeneity of variance are satisfied (Field, 2018). The test provides sufficient statistical power (1-β = .95) to detect medium-to-large effects (d ≥ 0.50) with the current sample size (N=80).

3.6 Ethical Considerations

The study received approval from the institutional ethics review board and participating schools' administrations. All participants and legal guardians provided written informed consent. Students could withdraw without penalty at any time. Anonymity was maintained through alphanumeric coding; only the principal

investigator held the linking key. Following data collection, control group students received access to AI tools to ensure equitable educational opportunity. All procedures adhered to Malaysian Personal Data Protection Act guidelines and international research ethics standards.

4. RESULTS

4.1 Pretest Equivalence

Table 1 presents descriptive statistics for pretest MUET writing scores, establishing baseline comparability between experimental and control groups.

Table 1

Descriptive Statistics for Pretest MUET Writing Scores

Group	N	Mean	SD	SEM	95% CI
Experimental	40	58.23	3.47	0.55	[57.11, 59.35]
Control	40	59.00	3.20	0.51	[57.97, 60.03]

Independent samples t-test revealed no significant pretest difference between groups, t(78)=-1.02, p=.31, d=-0.23, indicating successful baseline equivalence.

PENERBIT PRESS
UNIVERSITI TEKNOLOGI MARA

The small effect size and overlapping confidence intervals further confirm comparability, validating the quasi-experimental design's internal validity.

4.2 Posttest Outcomes

Table 2 displays descriptive statistics for posttest MUET writing scores following the four-week intervention period.

Table 2

Descriptive Statistics for Posttest MUET Writing Scores

Group	N	Mean	SD	SEM	95% CI
Experimental	40	91.33	4.39	0.69	[89.92, 92.73]
Control	40	82.78	3.01	0.48	[81.81, 83.75]

The experimental group demonstrated substantially higher mean scores (M=91.33) compared to the control group (M=82.78), representing an 8.55-point difference. The experimental group's larger standard deviation (SD=4.39 vs. 3.01) suggests greater variability in AI feedback responsiveness, possibly reflecting individual differences in digital literacy or self-regulated learning capacity.

4.3 Inferential Statistics

Table 3 presents independent samples t-test results examining posttest score differences between groups.

Table 3

Independent Samples T-Test for Posttest MUET Writing Scores

Test	F	Sig.	t	df	Sig. (2-tailed)	Mean Diff	SED	95% CI
Levene's	3.64	.060						
Equal var. assumed			10.16	78	<.001***	8.55	0.84	[6.87, 10.23]
Equal var. not assumed			10.16	68.98	<.001***	8.55	0.84	[6.87, 10.23]

Note. ***p < .001

Levene's test confirmed homogeneity of variance, F(1,78)=3.64, p=.060, permitting use of equal variance assumption. Independent samples t-test revealed a statistically significant difference favoring the experimental group, t(78)=10.16, p<.001, two-tailed. The 95% confidence interval [6.87, 10.23] excluded zero, corroborating significance. Cohen's d was calculated as:

$d = (M_1 - M_2) / SD\ pooled = (91.33 - 82.78) / 3.76 =$ **2.27**

This represents a very large effect size (Cohen, 1988), indicating that AI-assisted feedback produced substantial practical improvements beyond statistical significance. The magnitude suggests that the average student in the experimental group outperformed approximately 98.8% of control group students (based on normal distribution percentile conversion).

4.4 Analysis by Writing Component

PENERBIT PRESS
UNIVERSITI TEKNOLOGI MARA

To examine intervention effects across specific writing dimensions, posttest scores were disaggregated by MUET rubric criteria (Table 4).

Table 4
Mean Posttest Scores by Writing Component (out of 25 points each)

Component	Experimental M (SD)	Control M (SD)	t	p	Cohen's d
Task Fulfillment	22.15 (2.21)	20.35 (1.88)	4.02	<.001	0.87
Organization	23.40 (1.95)	21.10 (1.75)	5.74	<.001	1.24
Language Use	23.05 (2.08)	20.85 (1.82)	5.20	<.001	1.12
Mechanics	22.73 (1.89)	20.48 (1.67)	5.77	<.001	1.26

All components demonstrated statistically significant improvements favoring the experimental group (all $p<.001$). Effect sizes ranged from medium (Task Fulfillment, $d=0.87$) to large (Organization and Mechanics, $d>1.20$), with Organization and Mechanics showing particularly strong effects. This pattern suggests AI tools' greatest impact on structural coherence and surface-level accuracy, consistent with AWE framework predictions.

4.5 Summary of Findings

Pretest scores confirmed baseline equivalence between groups ($p=.31$). Posttest scores revealed significant experimental group superiority ($p<.001$, $d=2.27$). All writing components improved significantly under AI-assisted feedback. Largest effects occurred for Organization and Mechanics dimensions. The null hypothesis was rejected; AI-assisted feedback significantly enhanced MUET writing performance

5. DISCUSSION

5.1 Interpretation of Findings

Findings offer strong evidence to suggest that AI- aided feedback improves MUET essay writing performance than teacher -base traditional feedback. The magnitude of the very large effect ($d = 2.27$) exceeds benchmarks for practical significance commonly accepted in educational interventions (Hattie, 2009). These results are consistent with recent meta-analyses on the effectiveness of AI for language learning (Fitria, 2021; Wang et al., 2023) and also extend past research by showing effects within high-stakes assessment settings.

The analysis at the component level uncovers more subtle patterns, as all dimensions increased with the greatest growth in Organization and Mechanics. This differential profile is consistent with theoretical predictions of the AWE model (Warschauer & Grimes, 2008), which claims that AI tools are particularly strong in analysing structural coherence and surface error-checking – exactly those areas where the largest gains were observed. In contrast, Task Fulfillment (content quality and argumentation) was the features having the lowest effect size -AI may

PENERBIT PRESS
UNIVERSITI TEKNOLOGI MARA

not be well suited to infer high-level rhetorical issues- on AI capabilities in assessing higher-rational elements, which accords with Makwana's (2025) results concerning AI's difficulty for measuring argumentation depth.

5.2 Comparison with Previous Research

These findings support and expand previous research. Escalante et al. (2023) found that college students who received substantive AI feedback showed significant improvement in organization and self-revision, which are consistent with the Organization results of this study. Similarly, Quratulain et al. (2025) identified higher levels of accuracy and organization amongst Pakistani undergraduates who used AI virtual assistants but expressed concern with over dependence which was not tested in the present study and also merits examination.

It is worth mentioning that the effect size (d=2.27) of this study is much larger than other research findings. For instance, Shang (2024) reported moderate effects (d=0.62) on EFL writing proficiency for computer-assisted corrective feedback, whereas Taskiran et al. (2024) found small-to-medium effects (d=0.45) for distance language learners. Some reasons for this difference might include:

Intervention Intensity The current research used multi-tool AI feedback (Grammarly + Hemingway + QuillBot) as opposed to single tool avenues

High-Stakes Context: The role of MUET preparation in motivating and engaging students in feedback use

Pretreatment Performance: Participant pre-test scores (M≈58/100) indicated substantial need for improvement.

Control of Conditions: The quasi-experimental approach with equivalent instruction provided a more rigorous test of feedback effects than observational investigations.

5.3 Theoretical Implications

Results illustrate the importance of promtness and specificity in Feedback Intervention Theory (Kluger & DeNisi, 1996). AI provided feedback within 24 hr (vs. the traditional teacher feedback of 48-72 hours in control condition) for students to conduct revisions while compositions were still rememberable. This time advantage may have reinforced feedback usage and transfer of learning.

Furthermore, findings from this study confirm that the AWE program's understanding of AI as supplemental scaffolding and not a substitute for teacher (Warschauer & Grimes, 2008). Mechanically, structurally, and organizationally, AI content was as effective in providing feedback as teachers, but at the level of discourse and rhetoric teachers still had something to offer. The best pedagogy might combine the two: AI for immediate technical feedback, teachers for higher-order conceptual development.

5.4 Practical Implications

For teachers: AI feedback tools are a support to your natural feedback cycle, giving you back time for creating more content, teaching critical thinking skills and spending that one on one time in conferences. Educators need professional development not just to use AI tools, but also to be taught how they can responsibly

PENERBIT PRESS
UNIVERSITI TEKNOLOGI MARA

integrate AI apps, establish norms and realize pedagogical goals. Hybrid approaches, integrating AI's scale with teacher-sourced context, provide the most promising paths to effective ESL writing instruction.

For Learners: Automated feedback supports self-regulated learning with instantaneous, non-evaluative error identification that promotes iterative revision. Students build metacognitive awareness by reasoning about AI suggestions, and deciding which suggestions help them to their rhetorical ends. Teachers, though, need to foster critical AI literacy and urge students to not blindly accept the suggestions of an algorithm (Quratulain et al., 2025).

For Institutions: Think about incorporating AI writing tools into your English curriculum, especially when it comes to high-stakes exam preparation such as MUET. Execution is dependent on supporting infrastructure (sufficient and good-quality internet access, availability of devices) as well as policy frameworks for academic integrity, privacy and equitable access. Pilot projects with continuous evaluation can also be used to make evidence-based scaling decisions.

Implications for Policy Makers: These findings imply that it may be relevant to include AI language literacy and writing instruction in personal formation of national English curricula. The Malaysian Ministry of Education could create guidelines for ethical AI in education systems, where such tools would enhance not subvert human instruction. Investments in teacher training, technological infrastructure, and research on long-term outcomes would contribute to sustainable implementation.

5.5 Mitigating the Limitations and Risks of AI

Despite evidence of AI success, there are a number of limitations which should be taken into account. First, AI writing tools show limited proficiency in evaluating the quality of argumentation, rhetoric and cultural suitability (Makwana, 2025). In Malaysian contexts, AI systems that were developed using datasets consisting mainly of Western English may fail to identify local discourse practices and culturally shaped strategies of argumentation. Teachers must be on the look-out for culturally biased recommendations.

Second, the undermining of academic integrity and undue dependence should be preempted. Further, students may "over" rely on AI correction tools for their editing needs without learning editorial independence or abuse these tools to create rather than edit content (Quratulain et al., 2025). Educators should create clear-use policies, framing AI as a feedback tool not content engine and develop assessments which evaluate both process (rough drafts, rewrites) and products.

Third, there is a question of ongoing algorithmic bias. Such AI systems can further maintain linguistic biases, favoring particular dialects and types of discourse while sanctioning others.. In multilingual environments, such as those in Malaysia, students are exposed to different language inputs and should not take AI feedback for granted but critically rather.

PENERBIT PRESS
UNIVERSITI TEKNOLOGI MARA

5.6 Study Limitations

There are several limitations to generalizability and interpretation. First, the short duration of this four-week intervention does not allow us to draw conclusions about retention over time or transfer to new writing environments. Longitudinal research that monitors students across academic years would provide insight as to whether AI-assisted gains remain past immediate post-intervention timeframes.

Second, the sample of two schools limits external validity. The participants had broadly similar academic profiles (comparable Lower 6 results); generalisations to higher or lower achievers would be inappropriate. This could be improved through replication across different institutional settings – urban/rural, public/private and resource-level specific.

Third, no motivational, emotional or attitudinal factors were investigated. It is also anticipated that student involvement, anxiety, and self-efficacy as well as technology acceptance are potential moderators of AI impacts, but they were not addressed in measures. Qualitative studies of students' experiences, preferences and perceptions would complement the research.

Fourth, the degree to which the intervention was implemented with experimental group members differed slightly. Although all students had the same access to AI tools, self-reported logs (not independently analyzed) indicated differences in usage with some using one or two of the three more frequently than others. In the future, learning analytics can be used to measure tool use and its relationship to student outcomes.Fifth, characteristics of internal controls in instructor condition (experience, feedback quality) were not systematically manipulated. Differences in traditional feedback quality might have affected the results as well; however random assignment should have balanced this among conditions.

5.7 Recommendations for Future Research

There are several research directions after findings and limitations:

Longitudinal studies: Is the enhancement that AI provides sustainable, sustained across both semesters and generalizable to real academic writing situations beyond test preparation?

Mixed-Methods Approach: Concurrently, collect observational data about student experiences and use of AI feedback (i.e., interviews, think-aloud protocols, focus groups) as well as decision-making when using such feedback to optimize instructional pathways and their perceptions regarding the pedagogical value of AI.

Comparative Tool Studies: Carefully compare AI writing assistants with differences (Grammarly vs. ChatGPT vs. AWE that is designed only for English learners) to investigate which features are mostly effective in the context of ESL learning.

Moderator Analysis Analyze personal differences (digital literacy, writing self-efficacy, proficiency levels) which moderate AI efficacy to facilitate personalized matching of interventions.

PENERBIT PRESS
UNIVERSITI TEKNOLOGI MARA

Integration Models: Test inclusion approaches (AI-only, teacher-only, nested AI-then teacher, parallel AI-and-teacher) and highlight combinations that are most effective.
Delayed Retention Testing: Introduce long-term memory tests weeks or months after the intervention.
Diverse Populations: Replicate with different educational levels (secondary/tertiary) and ability groups (struggling vs. advanced writers), and in diverse linguistic contexts (different L1 populations).
Cost-Effectiveness: Compare cost of AI implementation to outcomes to provide information about how institutions allocate resources.
Research on Teacher's Perspective: Study teachers attitudes toward the use of AI tools, the challenges of integration and pedagogical modifications.
Ethical and Critical AI Literacy: To create, assess and train generations of students to critically appraise recommendations from the AI system and understand its limitations as well as handle it ethically.

5.8 Reflection and criticism on AI in education

While the present study provides evidence of substantial learning gains, more general discussions about AI and education are in need. Is this AI-aided improvement a real development of skill or a merely superficial enhancement of performance? Do leaners internalize and appropriately use the linguistic tendencies for when they are corrected live by AI or does external scaffolding overrule the need to deep process? These questions tap into more deep-seated tensions that play out in behaviorist models of learning (that focus on correctness) and constructivist ones (which focus on struggle for knowledge).
In addition, embedding AI also presents equity challenges. Students who have consistent access to technology are able to take advantage of it relative to peers who have been marginalized digitally, thus augmenting existing inequalities. They need to guarantee equitable access and guard against creating a two-tiered system in which affluent students are taught with the help of AI while others are not.
Lastly, AI's growing capabilities require continued reflection on what is distinctive about human expertise. As the tools approach, or surpass some Finite Human Capabilities (like spotting mistakes and registering patterns), educators need to proclaim, around which irreplaceable human capabilities are these replacement-proof: Empathy.* Cultural Competency,* Moral reasoning* Creative Idea-generating*. The objective is not AI for efficiency's sake, but a thoughtful integration that maximizes human capability while being realistic about the limitations of technology.

CONCLUSION

This quasi-experimental research has yielded strong empirical evidence that AI-assisted feedback significantly improves MUET writing output amongst Form 6 students. The strong effect size (d=2.27) and significant scores in all writing areas show that AI is pedagogically valuable in ESL settings. Results indicate that when

PENERBIT PRESS
UNIVERSITI TEKNOLOGI MARA

used judiciously, based on clear pedagogical goals, teacher control and student instruction, AI can be a powerful supplement to traditional writing instruction.
But AI shouldn't be considered a panacea or replacement teacher. Rather, they are best when they leverage AI's strengths (immediacy, consistency, scalability) alongside the permanently-human abilities of educators to influence their students (rhetorical guidance and cultural competence; personal ways of motivating). As AI advances, future work should study long-term results, diverse deployments, students' experiences and ethical considerations which will enable technology to support truly humanistic educational objectives.The research also adds to an increasing body of empirical evidence demonstrating the impact of AI for language learning whilst noting a number of remaining issues that would benefit from further exploration. Longitudinal studies using qualitative approaches with different populations can further develop the picture of how AI might be most effectively leveraged to facilitate ESL learners' writing development. As Malaysia and other countries move forward toward embedding educational technology applications, evidence-based decision-making based on strong research will be important to optimize returns while managing risks.In the end, AI is not a rival to human instruction; rather it's an enhancer that can enable educators by making timely, personalized and scalable feedback possible. In so doing, through reflexively embracing technology into both our professional practices and pedagogy, ESL teachers can improve learning results while maintaining the inimitable "human touch" in teaching as a fundamentally relationaland transformative practice.

ACKNOWLEDGEMENTS

The completion of this work owes much to many individuals and institutions. I thank everyone who made this work possible.

REFERENCES

Abdul Razak, N., & Md Yunus, M. (2021). A systematic review of literature on the use of technology for writing skills. *Creative Education, 12*(7), 1505–1518. https://doi.org/10.4236/ce.2021.127115

Alharbi, S. (2019). Effect of teachers' written corrective feedback on Saudi EFL university students' writing achievements. *International Journal of Linguistics, 8*(5), 15–29. https://doi.org/10.5296/ijl.v8i5.10197

Baker, P., & Potts, A. (2013). 'Why do white people have thin lips?' Google and the perpetuation of stereotypes via auto-complete search forms. *Critical Discourse Studies, 10*(2), 187–204. https://doi.org/10.1080/17405904.2012.744320

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin.

Chen, X., Zou, D., Xie, H., Cheng, G., & Liu, C. (2024). Two decades of artificial intelligence in education: Contributors, collaborations, research topics, challenges, and future directions. *Educational Technology & Society, 27*(1), 28–47. https://doi.org/10.30191/ETS.202401_27(1).0003

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Dimitrov, D. M., & Rumrill, P. D. (2003). Pretest-posttest designs and measurement of change. *Work, 20*(2), 159–165.

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education, 20*, Article 57. https://doi.org/10.1186/s41239-023-00425-2

Ferris, D. R. (2021). *Treatment of error in second language student writing* (3rd ed.). University of Michigan Press.

Field, A. (2018). *Discovering statistics using IBM SPSS Statistics* (5th ed.). SAGE Publications.

Fitria, T. N. (2021). Grammarly as AI-powered English writing assistant: Students' alternative for writing English. *Metathesis: Journal of English Language, Literature, and Teaching, 5*(1), 65–78. https://doi.org/10.31002/metathesis.v5i1.3519

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.

Hyland, K. (2019). *Second language writing* (2nd ed.). Cambridge University Press.

Hyland, K., & Hyland, F. (2019). *Feedback in second language writing: Contexts and issues* (2nd ed.). Cambridge University Press.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Lam, R. (2022). Teacher assessment literacy development: A narrative inquiry of writing teachers' concurrent beliefs and practices. *TESOL Quarterly, 56*(2), 558–590. https://doi.org/10.1002/tesq.3088

Lee, I. (2020). Utility of focused/comprehensive written corrective feedback research: Building theory, research and practice links. *Journal of Second Language Writing, 49*, Article 100724. https://doi.org/10.1016/j.jslw.2020.100724

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson Education.

Makwana, V. (2025). A comparative analysis of AI-powered and teacher-led feedback: Investigating student perceptions and writing performance. *Journal of English Language Teaching, 67*(1), 3–12. https://journals.eltai.in/jelt/article/view/JELT670102

Malaysian Examinations Council. (2021). *Malaysian University English Test (MUET): Regulations, test specifications, and test format*. Malaysian Examinations Council.

Pratama, A., & Sulistiyo, U. (2024). A systematic review of artificial intelligence in enhancing English foreign learners' writing skill. *PPSDP International Journal of Education, 3*(2), 170–181. https://doi.org/10.59175/pijed.v3i2.299

Quratulain, Maqbool, S., & Bilal, S. (2025). The effectiveness of AI-powered writing assistants in enhancing essay writing skills at undergraduate level. *Journal for Social Science Archives, 3*(1), 845–855. https://doi.org/10.59075/jssa.v3i1.166

Rethinasamy, S., & Chuah, K. M. (2016). The Malaysian University English Test (MUET) and its use for placement purposes: A predictive validity study. *Electronic Journal of Foreign Language Teaching, 13*(1), 85–101.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Shang, H.-F. (2024). Effectiveness of automated corrective feedback on EFL learners' writing proficiency and perception. *Asia Pacific Journal of Education*. Advance online publication. https://doi.org/10.1080/02188791.2024.2347318

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Stevenson, M., & Phakiti, A. (2019). *The effects of computer-generated feedback on the quality of writing*. Assessing Writing, 19, 51–65. https://doi.org/10.1016/j.asw.2013.11.007

Taskiran, A., Yazici, M., & Aydin, I. E. (2024). Contribution of automated feedback to the English writing competence of distance foreign language learners. *E-Learning and Digital Media, 21*(3), 287–304. https://doi.org/10.1177/20427530221139579

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wang, Y., Liu, C., & Tu, Y. F. (2023). Factors affecting the adoption of AI-based applications in higher education. *Educational Technology & Society, 26*(4), 116–129. https://doi.org/10.30191/ETS.202310_26(4).0009

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal, 3*(1), 22–36. https://doi.org/10.1080/15544800701771580

Wilson, J., & Roscoe, R. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research, 58*(1), 87–125. https://doi.org/10.1177/0735633119830764

Zhang, Z. V., & Hyland, K. (2023). Fostering student engagement with feedback: An integrated approach. *Assessing Writing, 55*, Article 100658. https://doi.org/10.1016/j.asw.2022.100658