AI-Assisted Academic Writing: A Quasi-Experimental Study Among Grade 10 Students

Marissa M. German

marissa.german@deped.gov.ph
Palo National High School, 6501, Philippines

Abstract

The continued difficulties in writing among high school-age Filipino students are also affected by the very large classroom sizes found in many Philippine Public Schools. This quasi-experiment evaluated the potential of AI-based writing tools to aid in the development of a student's academic writing skills. Over a four-week period, 50 students in two intact classrooms were divided into an Experimental Group that used the AI-based writing tool Grammarly for grammar checking and the AI-based brainstorming tool ChatGPT when completing essays, and a Control Group that received standard instructional techniques. Prior to treatment, there existed significant differences in the pretest writing skill abilities of the two intact groups (Experimental Mean = 18.28, Standard Deviation = 1.86; Control Mean = 12.62, Standard Deviation = 4.12; $t(48) = 8.06$, $p < .001$, Cohen's D = 1.60). To address these baseline ability differences, Analysis of Covariance was performed with the prior test as the covariate. Following the use of analysis of covariance to adjust for the baseline ability differences, results indicated that the Experimental Group significantly exceeded the Control Group on the posttest ($F(1, 47) = 11.05$, $p = .002$, Partial Eta Squared = .190; Adjusted Mean for Experimental Group = 18.10; Adjusted Mean for Control Group = 16.56). Furthermore, the Experimental Group demonstrated substantial within-group growth (Cohen's D = 0.91). On the Perceptions Survey, ratings from the Control Group indicated that they believed AI-based writing tools were more useful, easier to use, and more satisfying than those from the Experimental Group – a result contrary to expectations. However, both groups expressed a desire to continue using AI-based writing tools. Ninety-six percent of respondents from the Experimental Group expressed interest in continuing to utilize AI-based writing tools, and ninety-two percent of respondents from the Control Group expressed similar interest. Overall, the results suggest that AI-based writing tools can serve as intelligent scaffolding within Vygotsky's Zone of Proximal Development and provide a replicable example for the Department of Education's "Education Center for AI Research" (E-CAIR) initiative.

Keywords: artificial intelligence; academic writing; Grade 10; Grammarly; ChatGPT; automated writing evaluation; Philippine education

1. Introduction

Writing is arguably one of the biggest problems faced by all Filipino secondary students. It is widely acknowledged that there are several issues with grammar, organization, and coherence (Barrot, 2023); however, in public school classrooms with 40 to 50 students, these issues are likely to be exacerbated by the difficulty of providing individualized feedback to every student. As a whole, across the Philippine education sector, AI is being used increasingly to help mitigate the structural problem of having too many students to teach. Most Filipino educators and administrators support the use of AI to improve instruction and administration and to increase the time available for research, but have concerns about its impact on academic integrity and

contextualization (Giray et al., 2024). Moreover, recent meta-synthesis work has identified common themes related to ethical use, AI literacy, and inclusivity as foundational elements to create a framework for the responsible integration of AI (Funa & Gabay, 2025). Thus, at this point, there is a significant opportunity for research on integrating AI into writing instruction. Writing instruction is the area where the teacher's workload will be most tightly bound.

Recent meta-analyses provide some grounds for optimism. For example, a meta-analysis conducted by Fleckenstein et al. (2023) found an average effect size of .55 (medium) for AWE feedback on writing performance. Additionally, they found that the largest effect sizes were for transfer tasks (.65) compared to revision tasks (.27). Zhai and Ma (2023) also reported a large effect size (.86) for overall writing quality. Furthermore, Santiago et al. (2023) surveyed students in the Philippines and found that 82.87% of them had a positive experience with AI writing tools. Finally, Barrot (2023, 2024) demonstrated the value of using ChatGPT as a supplemental resource for second-language writers when it was appropriately integrated into instruction. However, two gaps exist. No previous research has investigated the use of AI-based writing tools in an experimental design with Grade 10 students in a public school classroom in the Eastern Visayas — a region with distinct linguistic and socio-economic characteristics.

Moreover, although prior studies have primarily focused on writing outcomes, few have explored attitudes toward usefulness, ease of use, and satisfaction alongside performance measures, which are essential for determining whether integrating technology into instruction is feasible over time (Davis, 1989; Venkatesh & Davis, 2000).

The theoretical rationale for the potential benefits of AI tools for writing instruction relies heavily on Vygotsky's (1978) theory of the Zone of Proximal Development (ZPD) and the broader body of research on scaffolding (Wood et al., 1976; Puntambekar & Hübscher, 2005). Specifically, AI tools can serve as dynamic, individualized scaffolds that provide just-in-time feedback. These types of scaffolds bridge the gap between what a writer can produce independently and what he/she can produce with support — thus adapting to each individual learner in ways that static resources (e.g., worksheets, grammar guides) could never do. In terms of instructional feasibility in a typical Philippine public school classroom with 50+ students, providing real-time feedback on grammar and ideational content may be the only viable alternative to one-on-one teacher commentary. With average workloads consisting of 200-250 students per year, it would be unrealistic for public school English teachers to consistently provide written or oral feedback to each student individually. If AI tools could provide adequate feedback on grammatical and ideational content areas, then perhaps teachers could focus their time on developing higher-level thinking skills such as argumentation, rhetoric, and critical thinking. This proposed study tested this hypothesis in one school and provides a model for replication across the remaining 1,342 public schools in the Leyte Division.

2. Methods

2.1 Research Design

The research utilized a quasi-experimental pretest-posttest control group design as described by Creswell

& Creswell (2023). Due to constraints of randomizing students in an intact classroom environment, two grade ten classes were randomly selected to participate. The experimental condition (n=25) utilized Grammarly for grammar checking and ChatGPT for brainstorming and outlining. The control condition (n=25), on the other hand, relied upon conventional methods of instruction (i.e., teacher feedback, peer review, dictionaries). Both groups were instructed by the same educator, thereby controlling for instructional effect. The time frame for the study included: Week one, introduction/orientation, student consent, pre-test, and AI tool training for the experimental group. Weeks two through three consisted of submitting two essays, one using the active treatment and one using the passive treatment. Week four consisted of the post-test and survey administration. The time spent in the active treatment portion of the study was approximately 14 of 28 days.

2.2 Setting

The research was conducted at a public high school in Palo, Leyte, that serves approximately 4,467 students. The school has a fully operational computer lab equipped with forty (40) functioning computers as well as an internet connection with speeds of twenty-five megabits per second. More than two-thirds of students, or about 68%, are from low-income families. As such, the school's physical resources far exceed those available in typical Philippine public schools; therefore, the results should be generalized only to other settings with similar resources.

2.3 Participants

The data were collected from 50 tenth-grade students aged 15 to 17. The students participated in two different groups. Eligibility for inclusion required that participants be currently enrolled as tenth graders, have parental permission and/or consent, and have basic computer skills. In addition, participants had to attend class at least 85% of the time during the study. Participants who were diagnosed with a learning disability that would require an individualized intervention program, and those with previous experience using artificial intelligence writing tools extensively, were excluded. All three demographics (gender, age, and family income) were comparable among all three groups. However, one demographic profile differed significantly: 60% of students who received the experimental treatment reported having previously earned an A (90 or higher) in their English classes. None of the responding control students indicated this level of achievement. This baseline skill difference is explicitly stated in Table 1 and analyzed in the analysis section (§2.6) using Analysis of Covariance (ANCOVA).

Table 1. Demographic characteristics: Experimental vs. Control Group (n = 25 each)

Variable / Category	Exp n	Exp %	Ctrl n	Ctrl %	Note
Grade Level: Grade 10	25	100.0	25	100.0	Both groups
Gender: Male	11	44.0	11	44.0	Identical
Gender: Female	14	56.0	14	56.0	
Age: 15	13	52.0	13	52.0	Similar
Age: 16	9	36.0	11	44.0	
Age: 17	3	12.0	1	4.0	Exp slightly older

Variable / Category	Exp n	Exp %	Ctrl n	Ctrl %	Note
Family Income: Low	6	24.0	5	20.0	Similar SES
Family Income: Middle	16	64.0	15	60.0	
Family Income: No Answer	3	12.0	5	20.0	
Prev. English: 89 Below	10	40.0	19	100.0*	Key difference
Prev. English: 90 Above	15	60.0	0	0.0*	

Note. Percentages are based on respondents' previous grade level in English (n = 25 experimental; n = 19 control). As demonstrated in section 3.2, the baseline difference in initial ability was supported by differences in pre-test performance.

2.4 Instruments

Academic Writing Rubric. An analytical rubric rated students along five criteria: organization, content, mechanics, language usage, and overall effectiveness. Each criterion was rated on a 5-point scale, with a maximum possible score of 25. Three high school English teachers validated the rubric. Pilot testing of the rubric was conducted with 20 essays from Grade 10 students to assess inter-rater reliability, yielding an ICC of .85.

Perception Survey of AI Tools. Perceived usefulness (five items) and ease of use (five items) were both measured using a five-point Likert scale (one = strongly disagree to five = strongly agree). Satisfaction with the tool was also measured via a five-item Likert scale. In addition, there were sections to assess how often the tool was used, its academic uses, the perceived benefits/drawbacks of using the tool, and whether the user would consider using similar technology in the future. The means of the data were categorized into four categories: Very High (>4.2- >5.0), High (3.4-4.2), moderate (2.6-3.4) low (<2.6- <3.4), very low (<1. 8-<2.6); and very low (<1.0- <1.8).

2.5 Procedure

Week 1 had two different orientations, one for each section. The Researcher presented both the traditional method of composition as well as the new technology, and handed out consent forms, which gave the students three (3) days to complete and submit. Students who returned their consent form(s) took a sixty (60)-minute pretest on computers in the computer lab, utilizing Microsoft Word with Internet disabled. They wrote a 250-300-word argumentative essay entitled "Should we be able to have our cell phones on in school?" On day three, the Experimental Group received a sixty (60)-minute hands-on orientation to Grammarly and ChatGPT. Each student in the Experimental Group also signed an Academic Integrity Agreement stating that although AI can assist in writing, it is not a replacement for writing.

For weeks 2 and 3, both groups wrote and submitted two (2) essays: an argumentative essay and an expository essay aligned with the Most Essential Learning Competency guidelines set forth by the Department of Education (DepEd). During this time, the Experimental Group used AI tools to draft and revise their essays. In addition to AI tools, the Experimental Group maintained brief reflection logs detailing how they utilized the tools. The

Control Group used peer reviews and dictionaries to aid in completing their essays. At week four, the post-test occurred. It consisted of a parallel 250–300-word argumentative essay about mental health that students wrote without AI. This measured what the students learned internally from prior experience with technology. Both groups completed a perception survey after completing the post-test. All essays were anonymous, assigned coded numbers, and graded by two trained raters who were unaware of the group designation or time frame during grading.

2.6 Data Analysis

The use of descriptive statistics provided an overview of demographic characteristics, writing scores, and survey responses. An independent-samples t-test was utilized to compare pretest scores. Normality was assessed using Shapiro-Wilk tests, while homogeneity of variance was assessed through Levene's tests. Since intact classes exhibited significant differences in writing abilities at baseline (See Section 3.2), the primary method for analyzing treatment effects was Analysis of Covariance (ANCOVA). Posttest served as the dependent variable, with Group as the fixed factor and Pretest as the covariate. Homogeneity of Regression Slopes was tested and met. To assess effect size, Cohen's d was calculated to determine within-group changes, and Partial Eta Squared (η^2) was calculated to determine effects from ANCOVA analyses. According to Cohen (1988), Partial Eta Squared values can be classified as follows: Small (.01), Medium (.06), or Large (.14); and, according to Cohen (1988), Cohen's d can be classified as follows: Small (0.20), Medium (0.50), and Large (0.80).

Analysis of survey data included descriptive statistics, one-sample t-tests comparing each mean score to a neutral midpoint of 3.0, and independent-samples t-tests comparing means between groups. For each subscale level, t-tests were conducted on item means (n=8) rather than on individual participants' subscale means; this limits statistical power, as discussed in Section 4.7. All tests used $\alpha = .05$.

2.7 Ethics

This research complied with the Data Privacy Act of 2012 (Republic of the Philippines, 2012) and DepEd's Research Guidelines. Ethical approval was given by both the Schools Division Office of Leyte and the school administration. Written parental permission for all participants, along with written student assent, after being fully briefed on the purpose, process, risk, and benefit. All participation in this research is voluntary. Participants can withdraw at any time without consequences or penalties. Participant identities are protected using coded names. All data collected during this research will be stored on a password-protected computer. Upon completion of the research, the control group will receive an equal opportunity to participate in training regarding the use of AI technology.

3. Results

3.1 Sample Characteristics

All 50 students completed both the pre- and post-tests. Table 1 shows that, as expected by randomization, the two groups had similar levels of demographics (gender, age, and income); however, they differed with respect to their past English grade: 60 percent of the experimental group reported a previous grade of 90 or higher, while no student from the responding control group did so. Section 3.2 quantitatively demonstrates how much greater the difference was in their prior scores.

3.2 Baseline Writing Performance

Pretest scores documented a substantial baseline gap. The experimental group scored markedly higher (M = 18.28, SD = 1.86) than the control group (M = 12.62, SD = 4.12). An independent-samples t-test confirmed the difference, $t(48) = 8.06$, $p < .001$, with a very large effect (Cohen's $d = 1.60$). This finding mirrors the previous English-grade disparity in Table 1 and illustrates the central limitation of intact-class quasi-experimental designs: the groups were not equivalent before the intervention. We therefore relied on ANCOVA, with pretest as covariate, for all subsequent treatment-effect analyses.

Table 2. Pretest Descriptive Statistics and Independent-Samples t-Test

Group	n	M	SD	t(48)	p
Control	25	12.62	4.12	8.06	<.001
Experimental	25	18.28	1.86		

Note. Pretest is the total rubric score (theoretical range 5–25). Cohen's d = 1.60, indicating a very large baseline difference favouring the experimental group.

3.3 Treatment Effects on Writing Performance

Unadjusted posttest means were 19.88 (SD = 1.42) for the experimental group and 14.78 (SD = 3.36) for the control group. Because of the baseline disparity, the central question was not whether the experimental group ended higher—it began higher—but whether it ended higher than its baseline alone would predict. The ANCOVA showed that it did: $F(1, 47) = 11.05$, $p = .002$, partial $\eta^2 = .190$. Estimated marginal means, evaluated at the grand pretest mean (M = 15.45), were 18.10 for the experimental group and 16.56 for the control group—an adjusted difference of 1.54 points. The effect exceeds Cohen's (1988) threshold for a large effect (.14).

Table 3. ANCOVA results for Posttest Scores using pre-test as covariates

Source	df	F	p	Partial η²
Pretest (covariate)	1	—	—	—
Group (treatment)	1	11.05	.002	.190
Error	47	—	—	—

Note. This table presents the results of an ANCOVA of the total rubric post-test score, with "group" as the fixed independent variable and "pre-test" as the covariate. The results indicate a positive effect for the "treatment" after accounting for the differences in the two groups' baseline performance.

Table 4. Estimated Marginal (Adjusted) Means for Posttest Writing Score

Group	Unadjusted M	Unadjusted SD	Adjusted M	Adjusted SE
Experimental (n = 25)	19.88	1.42	18.10	0.32

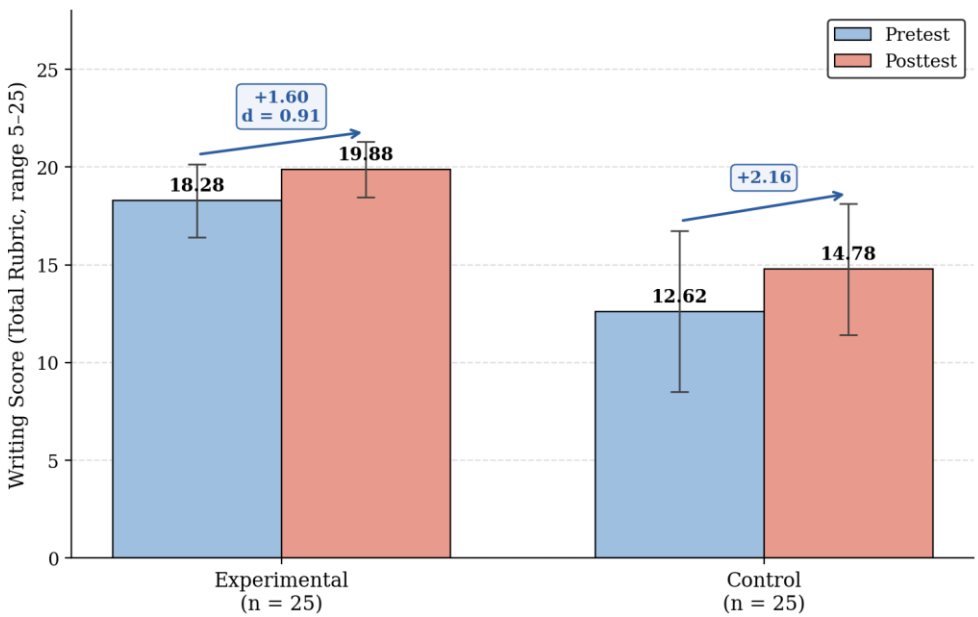
Group	Unadjusted M	Unadjusted SD	Adjusted M	Adjusted SE
Control (n = 25)	14.78	3.36	16.56	0.32
Adjusted Mean Difference	—	—	**1.54**	—

Note: The adjusted mean is calculated using the grand mean of the pretests (M = 15.45), and the adjusted standard error is derived from the pooled within-group MS-E to meet the assumption of homogeneous regression slopes.

3.4 Within-Group Gains

Both the experimental and control groups showed improvement from pretest to posttest. The experimental group demonstrated a 1.60-point increase from a pretest mean of 18.28 to a posttest mean of 19.88. This represented a strong internal (within-group) impact (d = .91). The control group increased by 2.16 points, from a pretest mean of 12.62 to a posttest mean of 14.78. At first glance, the control group appears to have made greater gains; however, this is due to its significantly lower initial mean rather than a higher degree of treatment responsiveness.

Figure 1. Writing Performance: Pretest vs. Posttest by Group



Note. Pretest non-equivalence: t(48) = 8.06, p < .001, d = 1.60. ANCOVA-adjusted treatment effect: F(1, 47) = 11.05, p = .002, partial η² = .190. Error bars represent ±1 SD.

Figure 1. Writing performance: pre-test vs. post-test for each group. The bar shows the mean total rubric scores (Range 5 – 25), while the error bars represent ±1 standard deviation. Pretest non-equivalence: t(48) = 8.06, p < .001, d = 1.60. ANCOVA-adjusted treatment effect: F(1, 47) = 11.05, p = .002, partial η^2 = .190.

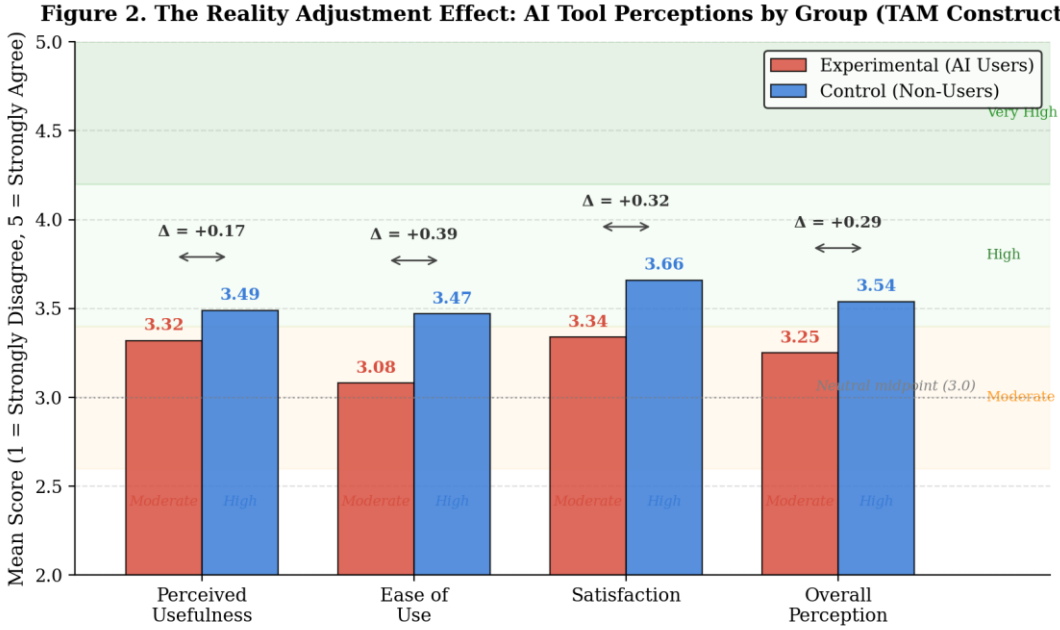
3.5 Comparative Student Perceptions of AI Tools

Students' perceptions of ChatGPT and Grammarly, as an application of the Technology Acceptance Model

(TAM), were measured using three constructs: perceived usefulness, ease of use, and satisfaction (each construct had five items). One-sample t-tests compared item means against the neutral midpoint (3.0); independent-samples t-tests compared groups. As noted in 2.6, subscale-level t-tests were computed on item means (df = 8) rather than participant-level means, which limits power.

The perception data revealed a counterintuitive pattern: the control group rated AI tools higher than the experimental group on all three TAM constructs (Figure 2). Perceived Usefulness was rated High by controls (M = 3.49) but only Moderate by the experimental group (M = 3.32). Ease of Use was rated High by controls (M = 3.47) and Moderate by the experimental group (M = 3.08). Satisfaction followed the same pattern (Ctrl M = 3.66, Exp M = 3.34). The overall composite score was lower for the experimental group (M = 3.25) than for the control group (M = 3.54). Group differences on individual subscales were not statistically significant at conventional thresholds (all p > .05), but the consistency of the pattern is theoretically meaningful and is examined in 4.2.

Figure 2. The Reality Adjustment Effect: AI Tool Perceptions by Group (TAM Constructs)



Note. The control group (non-users) rated AI tools higher than the experimental group (direct users) across all constructs — a pattern consistent with the "reality adjustment effect." Group differences on item-mean t-tests were not statistically significant (all p > .05).

Figure 2. The Reality Adjustment Effect: AI Tool Perceptions by Group (TAM Constructs). Students who did not have direct access to the AI tools (control group) had significantly higher ratings of utility, ease of use, and satisfaction with the AI tools compared to those who could access them directly (experimental group). This rating trend mirrors what has been referred to as the 'reality adjustment effect.' A comparison of independent-samples t-tests was conducted on the means of each question to assess statistical differences between groups. There were no statistically significant differences among these items: Perceived Utility t=−1.11, p=.301; Ease of Use t=−2.23, p=.056; Satisfaction t=−2.17, p=.062.

3.5.1 Perceived Usefulness

The mean score for perceived usefulness for the experimental group (M=3.32, SD=.21) scored moderately

high but statistically higher than the neutral midpoint of the scale (t [4] =3.38, p =.028). The most highly rated item by students in the experimental group was a1 regarding task effectiveness (M = 3.60), and the least highly rated item was a4 concerning time saving during revisions (M = 3.04). Four out of the five items were rated higher by students in the control group than those in the experimental group; however, both groups showed nearly identical ratings for a5 the perceived overall usefulness of the system. The independent-samples t-test was non-significant, t (8) = −1.11, p = .301.

3.5.2 Ease of Use

Of all the sub-scales measured by this survey, Ease of Use presented the biggest differences. The Experimental Group had a Mean (M = 3.08; SD = 0.28), which is the only subscale in both groups to have an average score greater than or equal to the neutral middle point, t(4) = 0.64, p = .556. In addition, as noted earlier, Item B5—Using AI Without Technical Difficulty—had the lowest rating in the entire survey with an Average Rating (M = 2.84). The Control Group rated "ease" considerably higher (M = 3.47) and created a larger differential (Δ = −0.39) than for any other item. Differences at the level of individual items were particularly large for Items B1 (Navigation Ease: Ctrl = 3.8 vs. Exp = 2.96) and B5 (Ctrl = 3.56 vs. Exp = 2.84). The between-group t-test approached but did not reach significance, t(8) = −2.23, p = .056.

3.5.3 Satisfaction

The Experimental Group's Satisfaction Mean (M= 3.34, S.D.= 0.25) was moderately high and was found to be statistically significant compared to the midpoint, T (4) = 3.09; P = .037. Of all five items, four were classified as moderately high, and C5 ("My overall experience with AI writing tools has been good") scored a high rating (M= 3.76), which had the highest rating for this category among the Experimental Group's survey. However, it should also be noted that C4 ("I would rather write using an AI than manually") scored the lowest in this category (M = 3.12), indicating that students who used these tools still do not feel they are better than doing things by hand. Overall, the Control Group rated their satisfaction higher (M = 3.66), with the largest differences observed in ratings for C1 (Satisfaction with feedback: Δ =- 0.56) and C2 (Willingness to Recommend: Δ =-0.52). Although there was once again a very close call regarding statistical significance, t (8) = -2.17; P=.062, it can be noted that the two groups only showed marginal differences when comparing themselves on C5 (Exp 3.76 vs. Ctrl 3.84); although there were statistically significant differences between the two groups for several other areas, both groups gave similar evaluations of their experiences.

3.6 Usage Patterns and Openness

3.6.1 Frequency of Use

The two groups used high-frequency AI in exactly the same way. Fifty-two percent of students in both groups used AI tools for academic work either "frequently" or "almost always." There was a slight difference within the groups as well. The Experimental Group used AI almost always (five times), whereas the Control Group used it frequently (12 times) and almost never (once). In terms of methodology, this does not matter. The Control Group was not naive about AI, as prior to and during the study, they used AI as much as the Experimental Group. Therefore, the treatment's effect on the students' writing could not have been due to their ability to access AI. The key that differentiated the two groups was not just the AI, but how they structured things around it.

3.6.2 Academic Purposes of Use

The overall pattern shows that across almost all activities, the control group used AI tools at a much higher rate than the experimental group. The greatest difference occurred in creating study materials (Experimental Group 20% vs. Control Group 60%) and in producing content (Experimental Group 16% vs. Control Group 40%)—the exact types of work the Experimental Group had been instructed by the Academic Integrity Agreement to avoid. The Experimental Group outperformed the Control Group only for gathering information (Exp. 68% vs. Ctrl. 72%; virtually identical) and brainstorming (Exp. 60% vs. Ctrl. 68%), the two subjects the treatment focused on. This trend suggests that structured pedagogy both reduced AI usage and changed the type of AI usage—specifically, encouraging idea formation rather than text generation.

3.6.3 Openness to Future Use

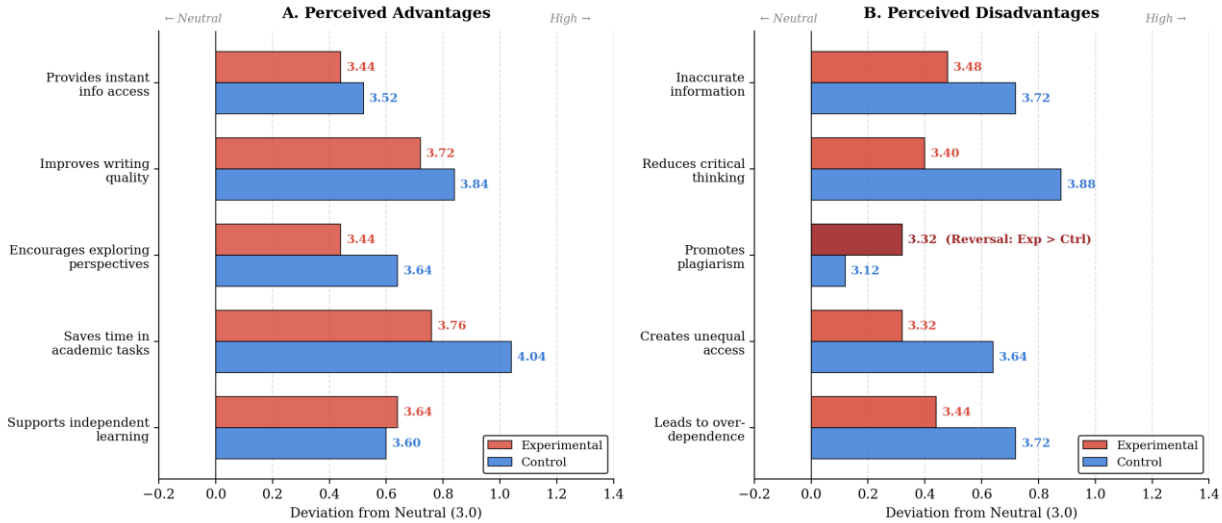
Openness to future use of AI was highly uniform in both groups. Experimental students were willing to continue using ChatGPT and Grammarly for all future academic writing and research. Experimental students (96%, or 24/25) and control students (92%, or 23/25) agreed they would continue to use these structured tools. Even among those who directly experienced bottlenecks due to friction with these structured tools, and despite the benefits, there was no change in their willingness to continue.

3.7 Perceived Advantages and Disadvantages

The two groups also showed a positive response to structured AI when asked about their perceived advantages. Experimental students rated each item on the advantages dimension positively (m = 3.60); control students also rated each item on the advantages dimension positively (m = 3.73). Experimental students rated only one advantage item significantly higher than control students did. Experimental students rated support for independent learning (3.64 vs. Ctrl M = 3.60) slightly higher than control students. Both groups most strongly believed that time-saving and improved accuracy are the principal advantages.

Theoretically, the greatest findings of interest in the section on Disadvantages came from the advantages section. Every disadvantage item was evaluated as less concerning by the experimental group than by the control group, except for one. The experimental group's concerns about a decrease in critical thinking, incorrect information, and unfair access to information were lower than those of the control group when using the tools (i.e., critical thinking: Exp 3.40 vs. Ctrl 3.88; $\Delta = -0.48$). The single exception was plagiarism. This time, the experimental group expressed greater concern about plagiarism (Exp 3.32 vs. Ctrl 3.12), as shown in Figure 3, highlighting the reversal in plagiarism concerns.

Figure 3. Perceived Advantages and Disadvantages of AI Tools by Group (Diverging from Neutral Midpoint)



Note. Bars show deviation from the neutral midpoint (3.0). Experimental students rated disadvantages lower (less concerned) than controls on 4 of 5 items, reflecting "informed reassurance." The single reversal on "promotes plagiarism" (darker bar) indicates heightened ethical awareness among AI users.

Figure 3. Perceived advantages and disadvantages of artificial intelligence tools by group (Diverging from Neutral Midpoint). Bars show deviation from Neutral Midpoint (3.0). Experimental students rated the disadvantages lower than controls on 4 of 5 items, reflecting "informed reassurance". The reversal on "promotes plagiarism" (darker bar in Panel b) indicates heightened ethical awareness among those who use artificial intelligence.

4. Discussion

4.1 Writing Performance

The main finding from this study is that AI-based writing assistance produced a significant difference relative to the control condition, even after controlling for a significant initial difference in students' writing abilities. The results obtained via analysis of covariance (ANCOVA), ($F(1, 47)=11.05$, $p=.002$, partial $\eta 2 =.190$), are congruent with the results of prior meta-analysis studies of automated writing evaluations (AWEs)—specifically Fleckenstein et al. (2023)'s estimate of an average effect of $g=0.55$ and Zhai & Ma's (2023) $g=0.86$—for providing empirical evidence supporting the applicability of AWE technology in the context of Philippines public schools. Additionally, the within-group effect size observed for the experimental group ($d=0.91$) is considered large, achieved in less than two weeks of active engagement with AI tools, indicating that these technologies can be used to create quantifiable improvements within a timeframe compatible with typical grading cycles. Using ANCOVA is important for understanding why. As a result of using an intact class design, there was a 1.6 SD baseline gap in writing abilities, or the type of gap that would make a post-test comparison meaningless if no statistical adjustments were applied. After statistically adjusting for pretest scores, the gap decreased from 5.10 raw points to 1.54 points; however, it remained. Therefore, it is reasonable to attribute the treatment effect to the specific intervention under study. In conjunction with equal rates of AI tool usage in both groups (high-frequent usage by approximately 52% of all students), this supports

one particular causal inference: the difference noted was due to how access to AI tools was framed rather than access itself (i.e., orientation, integrity frames, and task integration).

4.2 The Reality Adjustment Effect

The "Reality Adjustment Effect" is the most theoretically interesting result from our perception results. It appears that students who had experience using AI tools — meaning they encountered their limitations while completing graded essays under an integrity constraint — rated them lower than students who only encountered them casually.

Before using structured methods to examine how AI tools function, students usually perceived them through casual conversation, social media, and other forms of informal exploration. Those uses typically highlighted two features: how fast they were and how capable they seemed. However, when students actually use AI tools for assigned tasks — especially those requiring significant writing — they realize there are many more aspects of the tool they hadn't considered. For example, the Grammarly tool will catch some obvious errors in your grammar and spelling, but it will not help you develop a good structure to write a paper around. When you give ChatGPT a topic and ask it to generate ideas regarding that subject, what you get back from ChatGPT can sometimes require a great deal of editing and rewriting so that it meets the criteria of your assignment rubric or the standards of academic integrity that apply to the work. As a result of realizing such limits, students rate AI tools lower. But it isn't because their expectation of what AI tools could do has decreased -- it has simply become more realistic. This type of informed decision-making about technology is referred to as calibration (Long & Magerko, 2020) and is the type of knowledge that should be at the end of any program of study involving AI.

On the other hand, the significantly higher ratings given to AI tools by non-users (i.e., control subjects) represent an optimistic bias. In general, people have difficulty accepting that technologies they have not had to use in difficult situations may not perform as well as they expect.

4.3 Scaffolding Within the ZPD

This study demonstrates how Vygotsky's (1978) Zone of Proximal Development explains our findings. Grammarly and ChatGPT served as intelligent scaffolds — dynamic, personalized support systems — that helped students move from what they could accomplish on their own to what they could accomplish with support. It appears that three processes were involved.

First, AI-supported tools helped reduce cognitive load. Real-time grammar corrections allowed students to use their working memory to focus on other elements of the assignment, including organizing ideas, selecting relevant evidence, and creating an argument structure. Second, the AI tools offered student-by-student contingency. ChatGPT was more active when students had trouble coming up with an idea for an assignment; it became less active once students began generating independent ideas – a pattern consistent with the findings from Wood et al. (1976) on expert tutoring. Third, the rapid feedback cycle created opportunities for multiple cycles of drafting, reviewing, and revising. This process enabled students to create, identify errors, review, revise, and attempt again, all in one session, which aligns well with the "practice with feedback" model, as

documented in the writing research literature.

However, the experimental group's ease-of-use ratings present additional complexity. Although both groups reported being neutral about the overall ease of use (experimental M = 3.08), and specifically very low ratings on technical difficulties (experimental M = 2.84), these findings indicated that merely having access to a new tool does not result in effective scaffolding. Even though students received a full hour of orientation prior to using these tools, they still reported substantial usability difficulties. Therefore, the implications for E-CAIR are immediate: One-hour Week 1 orientations will never provide sufficient scaffolded practice for developing true competence and confidence with new tools.

4.4 Experience, Reassurance, and Specific Awareness

The results from the disadvantage section are very similar to those from the perception section. Students utilizing AI tools with an emphasis on integrity demonstrated lower concerns regarding three major risks (reduced critical thinking, inaccurate or misleading information, and unequal access) while demonstrating greater awareness/understanding of the one risk most closely aligned with writing ethically: plagiarism. This does not indicate a lack of awareness in the experimental group; rather, it reflects both increased confidence from experience and focused attention on one of these issues. In addition to recognizing that critical thinking is not being replaced but directed, students also recognize that they must redirect their critical thinking to identify inaccuracies and make rhetorical choices that no model can make for them. The increased concern for plagiarism demonstrates the same level of realistic understanding going in the opposite direction. We feel that this combination — less overall anxiety about using AI, and more focused awareness/understanding of specific ethical considerations — is exactly what we should see when we teach responsible pedagogical practices for AI.

4.5 Technology Acceptance and Behavioral Intention

The Technology Acceptance Model (TAM) by Davis (1989) and Venkatesh & Davis (2000) provides an appropriate framework for understanding the effects of perceived behavioral control. Both groups continued to report perceptions well above the neutral point on this scale and indicated extremely high levels of interest in using it again (96% in the experimental group and 92% in the control group). While behavioral intentions to use remained high, participants' subjective evaluations of specific usefulness and ease of use had some moderating effects. The C5 score (M=3.76) of the experimental group — the largest item in the experimental group's survey — illustrates another important finding. Although participants may have rated particular elements as being less useful individually, they generally had positive feelings about the overall experience. It seems that tools capable of producing a valuable outcome (e.g., writing improvements) can continue to receive user support, even though the moment-by-moment user experience was uneven.

4.6 Policy and Practical Implications

The research supports the Department of Education’s (DepEd) E-CAIR Policy Direction and empirically supports treating grammar correction tools as lower-risk when used appropriately. This research also provides an example of how to implement this within the 1342 public schools in Leyte Division: a fully equipped computer lab with access to reliable Internet is all the needed infrastructure; A four-week study period with a two-week

active intervention will allow it to be completed in time for teachers to grade their students at the end of each semester. Using the same teacher throughout the study eliminates the variable associated with different instructors. The free versions of Grammarly and ChatGPT have been found to be adequate for this study, and a group size of 25 will allow for the use of the current classroom structure.

4.7 Limitations and Future Directions

Several limitations qualify these findings. First, intact-class assignment produced a baseline gap of d = 1.60. ANCOVA adjusts for this statistically, but random assignment would provide stronger causal inference. Second, the active intervention spanned only two weeks. Fleckenstein et al. (2023) report that follow-up effects (g = 0.27) are smaller than immediate posttest effects (g = 0.57), so longitudinal designs are needed to assess retention. Third, the sample (n = 50) is modest; larger multi-school replications are warranted. Fourth, subscale-level perception analyses used df = 8 (item-level) rather than df = 48 (participant-level), substantially reducing power and likely contributing to the near-significant results for Ease of Use (p = .056) and Satisfaction (p = .062). Future studies should preserve participant-level data and report Cronbach's α for each TAM subscale. Fifth, while we report ANCOVA-adjusted means and the F-test, we did not report paired-samples Wilcoxon or paired-t inferential statistics for the within-group changes; these should be added in replications. Sixth, the study took place under direct supervision, and the findings may not generalize to unsupervised home use, where integrity risks are greater.

Future research should extend this work in several directions: longitudinal designs with delayed posttests to test retention; comparisons of AI-only feedback against combined AI–teacher feedback (Shi et al., 2025); proficiency-level moderation analyses; tests of generalization to narrative and descriptive writing; teacher-perspective studies on workload and pedagogy; and larger-sample replications of the reality adjustment effect to determine whether structured AI use reliably produces calibrated—rather than inflated—perceptions.

5. Conclusion

Grammarly and ChatGPT were used to create an AI-assisted writing instruction that positively impacted Grade 10 academic writing at a Philippine public school, even with a sizeable difference in writing abilities between the two intact groups. The ANCOVA-adjusted treatment effect was large (F (1, 47) = 11.05, p = .002, partial η^2 = .190), and within-group improvement in the experimental group was also large (d = 0.91).

Perception results changed how we define successful AI implementation. The students who actively used the tools gave them lower ratings than those with only informal access; however, their academic writing quality increased significantly and they demonstrated greater ethical use of the tools as well. Moreover, their willingness to continue using the tools was maintained at virtually the same level. Perception scores that are lower in this example are not indicative of a failed initiative -- they indicate calibrated knowledge. Students understood limitations of what AI tools could and could not accomplish, identified actual friction areas, and developed analytical evaluation techniques. In addition to developing positive attitudes toward technology, this type of learning will be more resilient over time than the enthusiasm generated from a lack of experience.

The most important implication of this study for practice is that it was the structure of the use of AI by the

students that produced differences in writing quality rather than their total amount of exposure to AI, suggesting that the key factor in successful integration of AI into teaching and learning will be "how" we provide access to AI, not simply "whether." In order to provide responsible and equitable access to AI, the three components of access (the availability of AI), ethical framing (the inclusion of some form of discussion regarding appropriate uses of technology), and pedagogical scaffolding (providing support for using technology appropriately within instructional settings) need to be integrated together.

Acknowledgements

The author would like to express gratitude to Dr. Mary Anne C. Sedanza and Mr. Elvin E. Morante of Leyte Normal University, who provided guidance on their research project as part of the PURPPLE Project; also, to the administration, faculty members, and 10th-grade students of Palo National High School for their assistance and participation during this research.

AI Acknowledgments

The author of this study wishes to thank Open AI's tools, Jenni and Kimi, for assistance with the organization, refinement, and development of certain aspects of this study. The author used AI as a resource to provide possible content structures, suggestions for improving clarity, and help ensure that all written components were consistent in style and tone. The final intellectual work, interpretation of findings, and conclusions remained entirely the researcher's own.

References

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, 57, 100745. https://doi.org/10.1016/j.asw.2023.100745

Barrot, J. S. (2024). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing*, 58, 100849. https://doi.org/10.1016/j.asw.2024.100849

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

Creswell, J. W., & Creswell, J. D. (2023). Research design: Qualitative, quantitative, and mixed methods approaches (6th ed.). SAGE Publications.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340. https://doi.org/10.2307/249008

Department of Education. (2025, February 20). DepEd launches AI center for education [Press release]. https://www.deped.gov.ph/2025/02/20/deped-launches-ai-center-for-education/

Department of Education & Microsoft. (2025, February 26). AI adoption in Philippine education [Press release]. https://news.microsoft.com/source/asia/2025/02/26/deped-and-microsoft-drive-ai-adoption-in-philippine-education/

Fleckenstein, J., Liebenow, L. W., & Meyer, J. (2023). Automated feedback and writing: A multi-level meta-analysis. *Frontiers in Artificial Intelligence, 6*, 1162454. https://doi.org/10.3389/frai.2023.1162454

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–16. https://doi.org/10.1145/3313831.3376727

Puntambekar, S., & Hübscher, R. (2005). Tools for scaffolding students in a complex learning environment. *Educational Psychologist*, *40*(1), 1–12. https://doi.org/10.1207/s15326985ep4001_1

Republic of the Philippines. (2012). Republic Act No. 10173: Data Privacy Act of 2012. Official Gazette. https://www.officialgazette.gov.ph/2012/08/15/republic-act-no-10173/

Santiago, C. S., et al. (2023). Utilization of writing assistance tools in the Philippines. *International Journal of Learning, Teaching and Educational Research*, *22*(11), 259–284.

Shi, Y., et al. (2025). ChatGPT as an automated writing evaluation tool. *Education and Information Technologies*, *30*, 1–25. https://doi.org/10.1007/s10639-025-13775-3

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model. *Management Science*, *46*(2), 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

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

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Zhai, X., & Ma, X. (2022). The effectiveness of automated writing evaluation on writing quality: A meta-analysis. *Journal of Educational Computing Research*, *61*(4), 875–900. https://doi.org/10.1177/07356331221127300