Self-Regulated Strategy Development and the Persuasive Writing of Grade 8 Learners

An Action Research

[Copy for blind review: the researchers' names, positions, and school are removed. The final copy lists them here.]

Conducted August to November 2025

Abstract

This action research tested whether Self-Regulated Strategy Development (SRSD), which teaches a planning strategy together with goal setting, self-monitoring, and self-instructions, would improve the persuasive essays of Grade 8 learners. Two Grade 8 sections of one English teacher in a public secondary school took part in Quarter 2 of School Year 2025–2026; a coin toss gave one section SRSD with the STOP and DARE strategy and kept the other on the usual process-writing lessons for the same 14 periods. Each learner wrote a persuasive essay before and after the unit, on two parallel prompts in counterbalanced order. Two trained raters, blind to section and occasion, scored all 170 essays of the 85 learners analyzed on a five-trait analytic rubric (5 to 20 points). Agreement was high for the totals, ICC(A,2) = .94, 95% CI [.91, .96], and acceptable for each trait, quadratic weighted κ = .63 to .75. An analysis of covariance with the pretest and the prompt order as covariates found that the SRSD section's posttest total was 2.17 points higher than the comparison section's, 95% CI [1.21, 3.13], $F(1, 81) = 20.27$, $p < .001$, Hedges' $g = 0.80$. After Holm's correction, the sections differed on three traits, the largest difference being in answering the other side's argument, 1.02 points of 4; the difference in language and conventions was not significant. The result held with each rater's scores alone. SRSD is recommended for Grade 8 persuasive writing, with conventions taught within its revising step.

Keywords: self-regulated strategy development, persuasive writing, analytic rubric, inter-rater reliability, analysis of covariance

Acknowledgment

The researchers thank the school head, who approved the study and kept the two sections' English periods unchanged for the quarter; the two English teachers who scored all 170 essays, and the teacher from a neighboring school who scored the traits on which they disagreed; the English department head, who observed the lessons; the Grade 9 learners and their parents, who allowed their practice essays to be used in training the raters; the Grade 8 advisers, who gave out and collected the consent forms; and the 91 Grade 8 learners and their parents. The study was funded by the researchers.

Self-Regulated Strategy Development and the Persuasive Writing of Grade 8 Learners

Context and Rationale

Many Filipino learners leave the primary grades unable to put an argument in writing. In the Southeast Asia Primary Learning Metrics of 2019, which tested Grade 5 learners' writing in English in the Philippines, 45% wrote at Band 1 or below, described as "Limited ability to present ideas in writing," and more than 70% were in the three lowest bands (UNICEF & SEAMEO, 2020). A persuasive essay asks more of a writer than any of those bands: a position, reasons that support it, an answer to the other side, and an order that holds them together.

Our Grade 8 learners showed the same need. When the first researcher taught Grade 8 in School Year 2024–2025, most persuasive essays stated a position only in the last sentence, gave reasons without explaining them, and ignored the other side. The pretest of this study found the same in both sections: on the Counterclaim trait, both raters scored 39 of the 85 pretest essays at 1, "No mention of the other side," and the two sections' pretest totals averaged 10.19 and 10.09 of 20. The MATATAG curriculum asks Grade 8 learners in the second quarter to "Analyze supporting evidence to validate assertions and counterclaims," to "Determine one's thesis as the central idea of the paper," and to "Compose the informational text based on the chosen text type," a persuasive text (Department of Education [DepEd], 2023, EN8INF-II-6, EN8INF-II-16, EN8INF-II-17).

Self-Regulated Strategy Development (SRSD) is among the most studied ways of teaching writing. In a meta-analysis of writing instruction for adolescents, teaching writing strategies had an average effect size of 0.82, and studies of SRSD averaged 1.14 against 0.62 for other strategy instruction; the process writing approach, which our comparison section kept, averaged 0.32 (Graham & Perin, 2007). In the elementary grades, SRSD averaged 1.17 (Graham et al., 2012). In a trial close to our setting, eighth graders in six Portuguese middle schools were taught opinion essays by their own teachers once a week for three months; the classes taught with SRSD "made statistically greater gains in composition elements than the comparison students immediately after instruction and two months later" (Festas et al., 2015, p. 17).

Whether SRSD helps Grade 8 learners in a public school's regular English periods, taught by their own teacher with the quarter's competencies, is therefore a question for our own classrooms. A study of writing also has a problem of its own: a teacher who scores their own learners' essays, knowing which section and which occasion each came from, can see the gains they hope for. This study had every essay scored by two trained raters who knew neither. It addresses the Teaching and Learning theme of the Basic Education Research Agenda (DepEd, 2016).

Innovation, Intervention, and Strategy

The innovation was Self-Regulated Strategy Development, taught in the persuasive writing lessons of one Grade 8 section in Quarter 2. Everything else was kept the same in the two sections: the teacher, the quarter's

competencies and texts, the model essays read in class, the writing topics practiced, and the time. Both sections had two writing periods a week, on Tuesdays and Thursdays, from September 2 to October 16, 2025: 14 periods of 45 minutes. The other three English periods of each week followed the quarter's reading and analysis lessons in both sections.

The Strategy: STOP and DARE

STOP and DARE is a planning strategy for persuasive essays (De La Paz & Graham, 1997). STOP reminds the writer to "Suspend judgment, Take a side, Organize ideas, and Plan more as you write" (p. 207); DARE names the parts the essay must have: "Develop your topic sentence, Add supporting ideas, Reject at least one argument for the other side, and End with a conclusion" (p. 208). Before taking a side, the learners listed reasons for both sides on a planning sheet; then they chose a side, numbered their reasons in the order they would use them, and kept planning as they wrote. The strategy fits the quarter's competencies, which ask learners to determine a thesis and to analyze the evidence for assertions and counterclaims in the texts they read (DepEd, 2023, EN8INF-II-16, EN8INF-II-6).

The Six Stages and the Self-Regulation Procedures

SRSD teaches a writing strategy together with the procedures that help learners manage it: goal setting, self-monitoring and self-evaluation, self-instructions, and self-reinforcement (Harris et al., 2011). Six stages of instruction carry both: "Develop and activate background knowledge," "Discuss it," "Modeling," "Memorize it," "Support it," and "Independent performance" (Harris et al., 2011, pp. 195–197). SRSD instruction is "criterion

based rather than time based" (p. 195): a learner moves on when they can do what the stage teaches. In a class of 46 with 14 periods, the plan gave each stage its lessons, and the teacher applied the criteria within them.

The 14 lessons followed the stages. In lessons 1 and 2, the learners read model persuasive essays and found their parts, and each learner counted the parts in their own pretest essay and graphed them, the first use of self-monitoring. In lesson 3, the teacher planned and wrote an essay aloud with STOP and DARE, saying the self-statements she used ("What is my side? What would someone who disagrees say? How do I answer that?"), and the learners wrote their own self-statements. The learners memorized the mnemonic from lesson 2 onward and recited it at the start of lessons 2 to 6. In lessons 4 to 10, the class planned and wrote essays together, then in pairs, then alone with a planning sheet whose prompts were removed one at a time. In lessons 11 to 14, the learners planned and wrote alone, set a goal for each essay, counted and graphed its parts, and checked their essays against their goals. The teacher moved a learner back to guided practice when the learner's graph showed a missing part twice in a row.

The Comparison Section

The comparison section was taught the same persuasive writing competencies in the same 14 periods with the teacher's usual process-writing lessons: reading and discussing a model essay, brainstorming as a class, outlining on a graphic organizer, drafting, receiving the teacher's written feedback, and revising. It wrote on the same practice topics and read the same model essays. It was not taught a mnemonic, self-statements, goal setting, or

the graphing of essay parts. It received SRSD in Quarter 3, from November 4, 2025, with the opinion editorials of that quarter.

Strategy and Evaluation Plan

The design compared the two sections' posttest essays, adjusted for their pretest essays. The English department head observed 6 of the 14 SRSD lessons with a checklist of each lesson's planned components, and 4 of the comparison section's writing lessons for any SRSD element. Three criteria for success were set in the proposal: (a) at least 12 of the 14 SRSD lessons held, with at least 80% of the planned components seen in the observed lessons; (b) scoring reliable enough to analyze, an ICC(A,2) for the essay totals whose lower 95% confidence bound reached .75 and a quadratic weighted kappa of at least .60 on every trait, both before disagreements were resolved; and (c) a posttest difference in the SRSD section's favor, significant at the .05 level with the pretest held constant, in the same direction when each rater's scores were analyzed alone.

Action Research Questions

Main question: Does Self-Regulated Strategy Development improve the persuasive writing of Grade 8 learners?

1. What were the pretest and posttest rubric scores of the SRSD section and the comparison section, in total and on each trait?
2. How closely did the two blind raters agree, on each trait and on the total, before and after their disagreements were resolved?
3. Is there a significant difference between the two sections' posttest totals when their pretest totals and the prompt order are held constant?

4. On which of the five traits did the two sections differ, when each trait's pretest score and the prompt order are held constant?

Two null hypotheses were tested at the .05 level of significance:

$H_0$1: There is no significant difference between the adjusted posttest totals of the SRSD section and the comparison section.

$H_0$2: On each of the five traits, there is no significant difference between the adjusted posttest scores of the two sections.

Action Research Methods

Research Design

The study used a quasi-experimental pretest–posttest design with a nonequivalent comparison group (Shadish et al., 2002). The learners could not be assigned one by one, because a class is taught as a section. The researcher-teacher's two Grade 8 sections were assigned by a coin toss on August 22, 2025, witnessed by the English department head. With two sections, the toss decided which section received SRSD but could not make the sections equal, so the pretest was used to check how similar they were and the analysis of covariance to adjust for the difference that remained.

Participants and Other Sources of Data and Information

The participants were the learners of two Grade 8 sections of a public secondary school in School Year 2025–2026, both taught English by the first researcher. All 91 enrolled learners were invited; no sampling was done. Table 1 shows how many were analyzed.

Table 1

Learners Enrolled, Excluded, and Analyzed, by Section

Learners	SRSD section	Comparison section	Total
Enrolled	46	45	91
No parental consent	2	1	3
Transferred to another school in September	1	0	1
Absent from the posttest and its make-up	1	1	2
Analyzed	42	43	85

Note. Learners who were not analyzed were taught with their sections; only their essays were left out. Only the essays of learners who wrote both were typed and scored.

The learners analyzed were 24 girls and 18 boys in the SRSD section and 23 girls and 20 boys in the comparison section. English was taught 45 minutes a day, five days a week, one of the class programs the MATATAG curriculum allows (DepEd, 2024a, 2024b). On the beginning-of-year Philippine Informal Reading Inventory, 3 learners of the SRSD section and 5 of the comparison section read English at the Frustration level; from September they attended ARAL-Reading sessions after classes, which took no English period (DepEd, 2025).

Sample Size

Because whole sections took part, the number of learners was fixed. A sensitivity analysis for an analysis of covariance with two covariates, the pretest and the prompt order, in G*Power (Faul et al., 2007) showed that the 86 learners expected after consent and transfers give a power of .80 to detect an effect of $f = 0.31$ at $\alpha = .05$, a standardized difference of $d = 0.61$ in units of the adjusted standard deviation. With a pretest–posttest correlation of .50, assumed in the proposal, that is $d = 0.53$ in units of the posttest's standard deviation, against $d = 0.61$ for a comparison of posttests without the pretest.

The pooled within-section correlation found in the study, .58, gave the 85 learners analyzed a detectable effect of $d = 0.50$.

Other Sources

Two English teachers of the school, who taught Grades 9 and 10 and had not taught either section, scored the essays. An English teacher of a neighboring school scored the traits on which they disagreed. The English department head observed lessons. A Grade 9 class of the school wrote the practice essays used to train the raters.

Data Gathering Methods

The Two Prompts

The learners wrote a persuasive essay on one of two prompts, built to be parallel: the same kind of claim, a claim of policy about school life; the same task; the same audience; and the same length of instructions. Prompt A read: "Your school head is deciding whether every Grade 8 learner should be required to join at least one school club. Write an essay to persuade the school head. State your position, give your reasons, and answer at least one argument of people who disagree with you." Prompt B asked the same about requiring every Grade 8 learner to take care of a part of the school's vegetable garden for one quarter. Two English teachers who were not part of the study judged both topics equally familiar to Grade 8 learners. Because two prompts are never perfectly parallel, the order was counterbalanced: before the pretest, half of each section's class list, drawn by lot, was assigned prompt A for the pretest and B for the posttest (order AB), and the other half the reverse (order

BA). Of the learners analyzed, 20 of the SRSD section and 20 of the comparison section had order AB.

Both essays were written in one English period, in 40 minutes, with paper, a planning sheet, and no dictionary: the pretest on Thursday, August 28, and the posttest on Tuesday, October 21, 2025, with make-up sessions the next day. The essays did not count toward the learners' grades.

The Analytic Rubric

The essays were scored on an analytic rubric of five traits, each from 1 to 4, for a total of 5 to 20 (Appendix C). The traits follow the quarter's competencies: Claim, the writer's position on the question (EN8INF-II-16); Reasons and Evidence, and Counterclaim, the support for the position and the answer to the other side (EN8INF-II-17, building on EN8INF-II-6); Organization, the opening, the order of ideas, the transitions, and the ending (EN8INF-II-2, EN8INF-II-19); and Language and Conventions, sentences, word choice, grammar, spelling, and punctuation (EN8INF-II-20, EN8INF-II-21). Each score point has a descriptor; a 4 on Counterclaim, for example, reads "States an argument of the other side fairly and answers it with a reason or evidence," and a 2, "Mentions that others disagree but does not say what they argue, or states it without an answer." An analytic rubric was chosen over a single holistic score because the research on rubrics finds that reliable scoring "can be enhanced by the use of rubrics, especially if they are analytic, topic-specific, and complemented with exemplars and/or rater training" (Jonsson & Svingby, 2007, p. 130), and because a score for each trait shows where the writing changed.

Rater Training

The two raters were trained in one three-hour session on November 7, 2025. The anchor essays came from 34 practice essays that a Grade 9 class of the school had written in Quarter 1 on a third prompt of the same kind, with their parents' consent. The researchers chose 10 of them as anchors, essays that showed each score point of each trait, and the raters discussed the rubric trait by trait against the anchors. They then scored 12 other practice essays alone and compared: they agreed exactly on 48.3% of the 60 trait scores and within one point on 93.3%. After they discussed every disagreement against the descriptors and the anchors, they scored 12 more on November 10, agreeing exactly on 73.3% and within one point on 100.0%. The rule set in advance, at least 70% exact and 95% within one point, was met in the second round, and scoring began.

Blind Scoring

Every essay was typed word for word by the first and second researchers, keeping the learners' spelling, punctuation, and paragraphing, and the second researcher checked each typed essay against its paper copy. Typing removed the handwriting, which could tell a rater whose essay it was, and the paper itself, which could tell its date. Names of persons in the essays were replaced with [name]. The planning sheets were not typed, because the SRSD section's would have shown the mnemonic. Masking the writer's identity, keeping handwriting from swaying the score, and putting the papers in random order before scoring are among the recommendations of a review of writing assessment, which found that the legibility of handwriting, and

knowing a writer's gender, disability or ethnicity, both move the scores raters give (Graham et al., 2011, pp. 17, 21–22).

The second researcher, the school's research coordinator, gave each of the 170 essays a random four-digit code and kept the key, which linked codes to learners, sections, occasions, and prompts, in a password-protected file the raters never saw. The essays were shuffled and split into 10 batches of 17, each batch a mix of both sections, both occasions, and both prompts. Each rater received the batches in a different random order, and the essays within a batch in a different order. The raters scored independently from November 11 to 21, 2025, on a scoring sheet for each essay, and did not discuss any essay.

Resolving Disagreements

Where the two raters' scores on a trait differed by one point or not at all, the final score was their mean. Where they differed by two or more points, the trait went to a third rater, trained with the same anchors, who scored it without seeing the first two scores; the final score was the mean of the third rater's score and the closer of the first two, or the third rater's score when the two were equally close. The final scores, from 1 to 4 in steps of 0.5, were the data analyzed. The key was opened only after every final score was recorded, on November 26, 2025.

Inter-Rater Reliability

Agreement was computed on the raters' independent scores, before any resolution, because the resolved scores agree partly by construction. For each trait, we report exact agreement, agreement within one point, and Cohen's

(1968) weighted kappa with quadratic weights, which counts a disagreement of two points four times as heavily as one of one point; its 95% confidence interval came from 2,000 bootstrap resamples of the essays. For the totals, we report the intraclass correlation coefficient from a two-way random-effects model for absolute agreement (McGraw & Wong, 1996): ICC(A,1) for a single rater and ICC(A,2) for the mean of the two raters, the score the study analyzed. A random-effects model treats the raters as two of the trained raters the school could have used, and absolute agreement counts a rater's general severity as error, since a rater who scores every essay a point lower is not interchangeable with one who does not. Each ICC was judged by its 95% confidence interval against Koo and Li's (2016) guidelines: below .50 poor, .50 to .75 moderate, .75 to .90 good, and above .90 excellent. Exact agreement was reported beside the 70% often used as a guideline; agreement within one point says little on a four-point scale, where "it would be surprising to find agreement lower than 90%" (Stemler, 2004, p. 2). Quadratic weighted kappa is, apart from a term that shrinks as the sample grows, an intraclass correlation that counts the raters' differences in severity as error (Fleiss & Cohen, 1973). To check that the raters did not know the essays' occasion or section, each rater guessed both for 24 essays drawn by the second researcher, six from each section and occasion, after scoring was finished.

Implementation Records

The first researcher kept a log of each writing period and each learner's attendance in both sections. The English department head observed lessons 2, 4, 6, 8, 11, and 13 of the SRSD section with a checklist of each lesson's

planned components, 6 to 8 per lesson, and lessons 3, 7, 10, and 12 of the comparison section for any of four SRSD elements: the mnemonic, self-statements, goal setting, and the graphing of essay parts.

Storage

The paper essays, planning sheets, and consent forms were kept in a locked cabinet at the school. The typed essays, the scores, and the data set carry only codes; the key linking codes to names stayed with the second researcher, apart from the data.

Ethical Considerations

The school head approved the study in writing on August 15, 2025. Parents received a letter in English and Filipino that explained the purpose, the two essays, the scoring by teachers who would not know whose essays they read, the risks and benefits, and the right to withdraw, and signed a consent form; learners gave their assent in the English period on August 27, 2025, after the same explanation. Every learner received the same lessons and wrote the same essays whether or not they took part; only the essays of consenting learners were typed and scored. The essays did not count toward grades. A learner's essay is personal information when it can be traced to the learner (Republic Act No. 10173, 2012, Sec. 3(g)), so the essays were coded and typed before anyone outside the research team read them, and personal details in them were masked. The researchers agreed in advance that an essay disclosing harm to a learner would be referred to the school's Child Protection Committee (DepEd, 2012, Sec. 10); none did. Because the comparison section could benefit from SRSD, it received the strategy in Quarter 3. The study

followed the Research Management Guidelines (DepEd, 2017). The researchers' use of an artificial intelligence (AI) tool is declared in Appendix F (DepEd, 2026).

Data Analysis

For the first question, the totals and trait scores were summarized by section and occasion, and the sections' pretest totals were compared with Welch's *t* test. For the second, the agreement statistics above were computed for each trait and for the totals, before and after resolution; the raters' totals were compared with a paired *t* test to measure their severity; and the raters' guesses of section and occasion were compared with chance by a binomial test.

For the third, an analysis of covariance (ANCOVA) compared the sections' posttest totals with the pretest total and the prompt order (AB or BA) as covariates; the order covariate takes out any difference in the prompts' difficulty, which counterbalancing spreads evenly over the two sections. The adjusted difference is reported with its 95% confidence interval, *F*, *p*, and partial η^2, and as Hedges' *g*, the adjusted difference divided by the pooled standard deviation of the posttest, with the small-sample correction (What Works Clearinghouse [WWC], 2022). The WWC uses the unadjusted standard deviation even when the difference is adjusted, and judges two groups' pretests equivalent when they differ by 0.05 standard deviations or less; a difference between 0.05 and 0.25 must be adjusted for in the analysis (WWC, 2022, pp. 53, 162). The ANCOVA's assumptions were checked: equal slopes of posttest on pretest in the two sections (the section × pretest interaction),

normal residuals (Shapiro & Wilk, 1965), and the prompt order acting alike in both sections (the section × order interaction).

Six checks followed: the ANCOVA on each rater's totals alone, on the raters' mean before resolution, without the learners in ARAL-Reading, and with heteroskedasticity-consistent (HC3) standard errors; and a Mann–Whitney *U* test of the gains, which uses only their ranks, since rubric totals are counts of score points rather than measurements on an interval scale.

For the fourth question, the same ANCOVA was run for each trait, with that trait's pretest score as covariate, and the five *p* values were adjusted with Holm's (1979) step-down procedure so that the chance of any false rejection among the five stayed at .05.

The analyses were run in Python 3.12 with pandas, statsmodels (Seabold & Perktold, 2010), SciPy, and scikit-learn, free software; the Excel data set computes the agreement statistics, the ICC, the ANCOVA, and Holm's correction as formulas, and the notebook reproduces every result (Appendix A).

Discussion of Results and Reflection

Implementation

All 14 SRSD lessons were held; one was moved to the next day after a class suspension for rain, and the comparison section's lesson of that week was moved with it. The learners analyzed attended a mean of 13.3 of the 14 writing periods in the SRSD section (range 12–14) and 13.2 in the comparison section (range 12–14). In the six SRSD lessons observed, 37 of the 40 planned components were seen (92.5%); no observed lesson fell below 83.3%. None of

the four SRSD elements was seen in the 4 comparison lessons observed. The first criterion for success was met.

Question 1: Scores Before and After

The sections began at nearly the same level, and the SRSD section rose more (Table 2). Its mean total rose by 3.37 points, the comparison section's by 1.24. The pretest means differed by 0.10 points, $t(81.23) = 0.16$, $p = .870$, a difference of 0.04 standard deviations. The SRSD section's essays also grew longer, from 104 to 154 words on average, against 97 to 120 in the comparison section.

Table 2

Pretest and Posttest Scores by Section, in Total and by Trait

Score	SRSD pretest *M* (*SD*)	SRSD posttest *M* (*SD*)	Comparison pretest *M* (*SD*)	Comparison posttest *M* (*SD*)
Total (5–20)	10.19 (2.50)	13.56 (2.67)	10.09 (2.97)	11.34 (2.72)
Claim (1–4)	2.38 (0.62)	3.10 (0.67)	2.34 (0.71)	2.71 (0.68)
Reasons and evidence	2.10 (0.81)	2.80 (0.72)	2.03 (0.79)	2.29 (0.73)
Counterclaim	1.39 (0.46)	2.60 (0.75)	1.45 (0.49)	1.59 (0.58)
Organization	2.21 (0.75)	2.74 (0.67)	2.15 (0.74)	2.38 (0.76)
Language and conventions	2.11 (0.75)	2.33 (0.78)	2.12 (0.89)	2.36 (0.80)

Note. SRSD section $n = 42$; comparison section $n = 43$. Final scores: the mean of the two raters, or the resolved score where they differed by two or more points.

Question 2: The Raters' Agreement

The raters agreed exactly on 59.6% of the 850 trait scores and within one point on 98.7%; no two scores were three points apart (Table 3). Exact agreement fell short of the 70% guideline on every trait, from 53.5% to 67.6%; none of 21 studies of analytic scales in a review of writing assessment met it either (Graham et al., 2011, p. 24). Their quadratic weighted kappas ranged

from .63 to .75, and every trait met the criterion of .60. Their totals agreed closely, ICC(A,1) = .89, 95% CI [.84, .92], for a single rater, and ICC(A,2) = .94, 95% CI [.91, .96], for the mean of the two. The second criterion for success, which rested on the kappas and the ICC because they credit near misses, was met.

Table 3

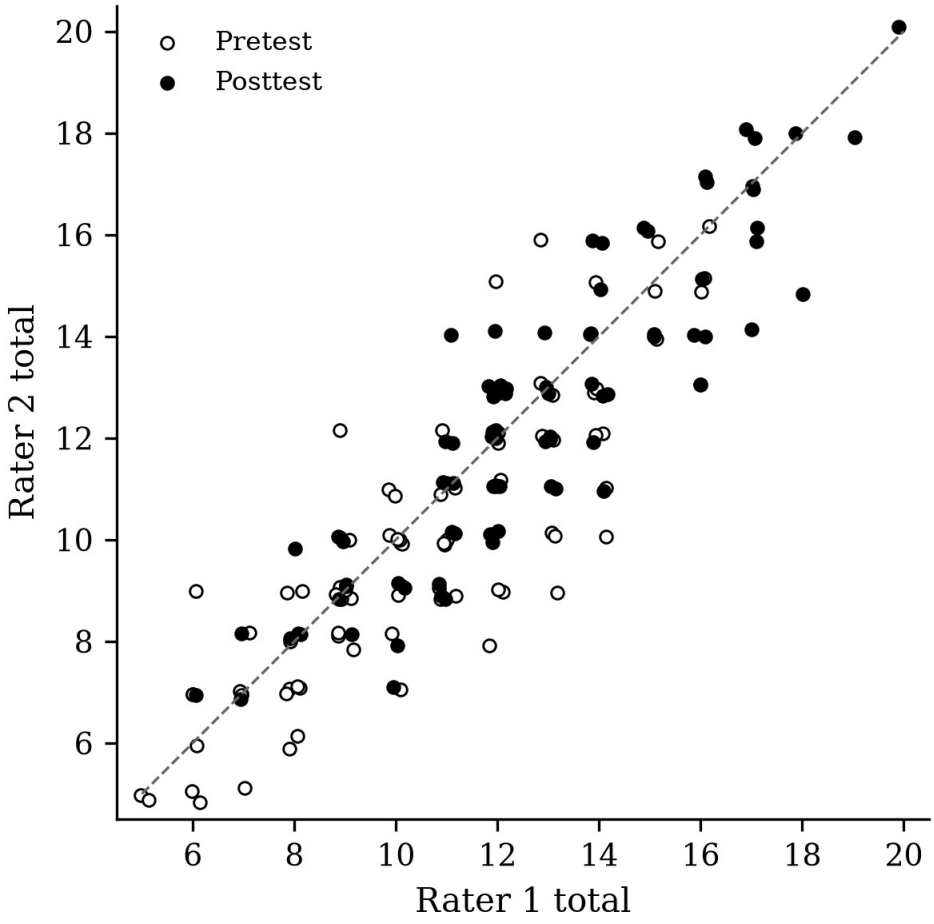
Agreement Between the Two Raters Before Resolution, by Trait

Trait	Exact agreement (%)	Within one point (%)	Quadratic weighted κ [95% CI]	κ, pretest essays	κ, posttest essays	Rater 1 minus rater 2
Claim	57.1	98.2	.63 [.54, .71]	.57	.60	−0.05
Reasons and evidence	62.9	100.0	.75 [.69, .81]	.73	.74	−0.06
Counterclaim	67.6	98.8	.73 [.63, .80]	.41	.78	0.11
Organization	57.1	98.2	.66 [.56, .73]	.70	.58	0.14
Language and conventions	53.5	98.2	.68 [.60, .74]	.65	.70	0.32

Note. 170 essays, each scored by both raters: 85 pretest and 85 posttest essays. κ = quadratic weighted kappa; its confidence interval is from 2,000 bootstrap resamples of the essays. The last column is the mean difference between the raters' scores, in points of 4.

Figure 1

The Two Raters' Totals for Each Essay



Note. Each point is one essay; points are jittered slightly so that essays with the same totals can be seen. The dashed line marks equal totals.

Two features of the agreement shaped the analysis. First, the raters differed in severity: Rater 1's totals averaged 0.46 points higher than Rater 2's, $t(169) = 4.22$, $p < .001$, most of it on Language and Conventions (0.32 points). Because both raters scored every essay of both sections, Rater 1's leniency entered every final score alike and could not favor either section; it is why the absolute-agreement ICC(A,1), .89, is a little below the consistency ICC(C,1), .90. Second, on the Counterclaim trait of the pretest essays, the

raters agreed exactly on 65.9% of essays, more than on any other trait, yet its kappa was the lowest of the pretest traits, .41. Both raters gave 39 of the 85 pretest essays a 1, and when most scores sit on one level, kappa expects much of the agreement by chance. Feinstein and Cicchetti (1990) called this a paradox of kappa: high agreement "can be drastically lowered by a substantial imbalance in the table's marginal totals" (p. 543). On the posttest essays, whose scores spread over all four levels, the trait's kappa was .78.

The raters differed by two or more points on 11 of the 850 trait scores (1.3%), in 11 essays, and the third rater scored those traits. Resolution changed the totals of 4 essays, by at most 1.0 point; in the other cases the third rater's score fell midway and equaled the first two raters' mean. After resolution, ICC(A,2) was .95, 95% CI [.92, .96]. Resolution narrows disagreement by design, which is why the reliability reported above is the one before it.

The raters could not tell which section an essay came from. After scoring, Rater 1 named the section of 13 and Rater 2 of 10 of the 24 essays drawn for the check, near the 12 that guessing would give (binomial p = .839 and .541). They named the occasion of 13 and 16 (p = .839 and .152).

Question 3: The Difference Between the Sections

With the pretest total and the prompt order held constant, the SRSD section's posttest total was 2.17 points higher than the comparison section's, 95% CI [1.21, 3.13], $F(1, 81) = 20.27$, $p < .001$, partial $\eta^2 = .20$, Hedges' g = 0.80 (Table 4 and Figure 2). The adjusted means were 13.53 and 11.36 of 20. $H_0 1$ is rejected.

Table 4

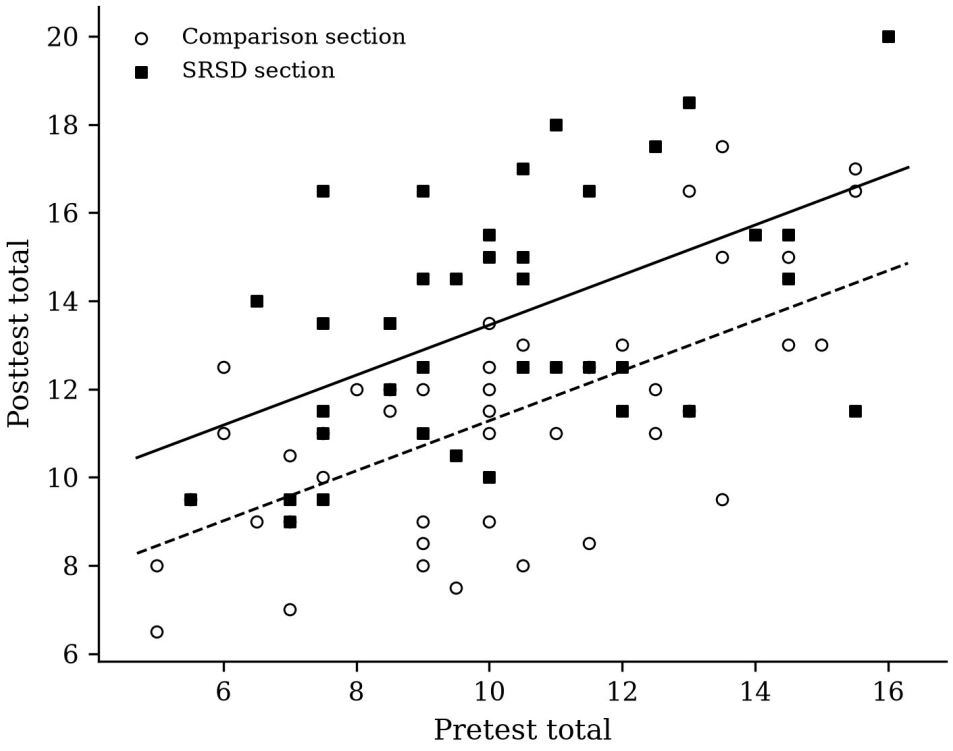
Analysis of Covariance of the Posttest Total

Term	*b*	*SE*	*t*	*p*	95% CI
Intercept	5.47	0.99			[3.50, 7.45]
Pretest total	0.57	0.09	6.39	< .001	[0.39, 0.74]
Prompt order (BA vs. AB)	0.26	0.48	0.53	.598	[−0.71, 1.22]
Section (SRSD vs. comparison)	2.17	0.48	4.50	< .001	[1.21, 3.13]

Note. $N = 85$; $df = 81$. $R^2 = .43$. Ordinary least squares with the final scores. Adjusted means at the mean pretest total and the mean prompt order: SRSD section 13.53, comparison section 11.36.

Figure 2

Posttest Total Against Pretest Total, by Section, With the Fitted ANCOVA Lines



Note. Each point is one learner. The two parallel lines are the ANCOVA's fitted values at the mean prompt order; the vertical distance between them is the adjusted difference.

The ANCOVA's assumptions held. The slopes of posttest on pretest did not differ significantly between the sections, $F(1, 80) = 0.26$, $p = .608$; the residuals showed no departure from normality, Shapiro–Wilk $W = .99$, $p =$

.495; and the prompt order did not act differently in the two sections, $F(1, 80) = 0.00$, $p = .960$. The two prompts proved close to parallel: within learners, prompt B scored 0.12 points lower than prompt A, and the order covariate's coefficient was 0.26, $p = .598$.

The checks agreed with the main result (Table 5). Each rater's scores alone gave a difference of about two points, as did the raters' mean before resolution and the analysis without the 8 learners in ARAL-Reading. The Mann–Whitney test of the gains, which assumes only that they can be ranked, gave the same conclusion, $U = 482.5$, $p < .001$, rank-biserial $r = .47$; the median gain was 3.75 points in the SRSD section and 1.50 in the comparison section. The third criterion for success was met.

Table 5

The Adjusted Difference Under Five Checks

Analysis	*n*	Difference (points)	95% CI	*p*
Main analysis: final scores	85	2.17	[1.21, 3.13]	< .001
Rater 1's totals alone	85	2.01	[0.98, 3.04]	< .001
Rater 2's totals alone	85	2.45	[1.47, 3.43]	< .001
Mean of the raters before resolution	85	2.22	[1.28, 3.17]	< .001
Without the learners in ARAL-Reading	77	2.13	[1.12, 3.14]	< .001
HC3 standard errors	85	2.17	[1.20, 3.14]	< .001

Note. Each row is the ANCOVA of the posttest total on the pretest total, the prompt order, and the section.

Question 4: Trait by Trait

The sections differed most on Counterclaim, by 1.02 points of 4, and also on Reasons and Evidence and on Claim, after Holm's correction (Table 6 and Figure 3). The difference on Organization, 0.33 points, was significant at .05 alone, $p = .027$, but not after the correction, $p = .054$. On Language and

Conventions the sections did not differ significantly, −0.02 points, 95% CI [−0.29, 0.25]. $H_0 2$ is rejected for Claim, Reasons and Evidence, and Counterclaim, and not rejected for Organization and Language and Conventions.

Table 6

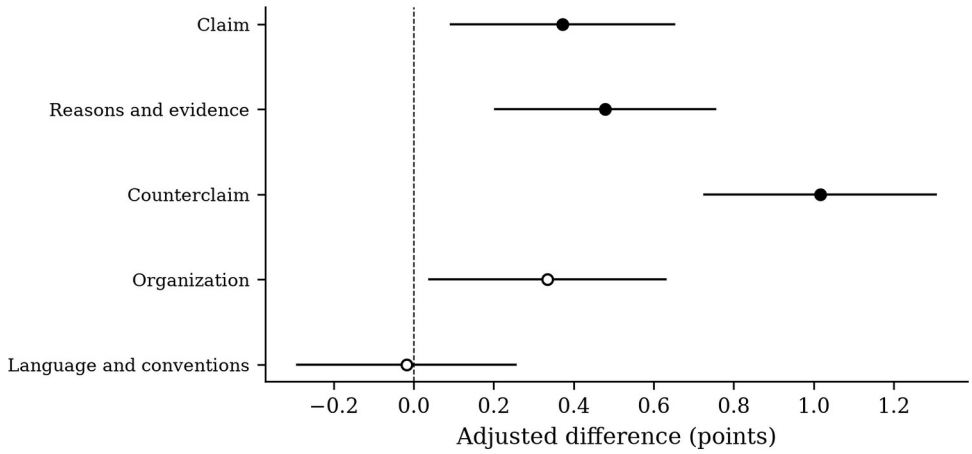
Analysis of Covariance of Each Trait's Posttest Score

Trait	Adjusted difference	95% CI	*F*(1, 81)	*p*	Holm-adjusted *p*	Partial η²
Claim	0.37	[0.09, 0.65]	7.05	.010	.029	.08
Reasons and evidence	0.48	[0.20, 0.75]	12.00	< .001	.003	.13
Counterclaim	1.02	[0.73, 1.30]	48.88	< .001	< .001	.38
Organization	0.33	[0.04, 0.63]	5.07	.027	.054	.06
Language and conventions	−0.02	[−0.29, 0.25]	0.02	.893	.893	.00

Note. Each trait is scored 1 to 4. Each row is an ANCOVA of the trait's posttest score on its pretest score, the prompt order, and the section. The confidence intervals are not adjusted for the five tests; the Holm-adjusted *p* values are.

Figure 3

Adjusted Differences Between the Sections on the Five Traits



Note. Points are adjusted differences (SRSD minus comparison) with 95% confidence intervals. Filled points are significant after Holm's correction.

Reflection

What we found. SRSD raised the persuasive essays of the Grade 8 section that received it by 2.17 points of 20 over the comparison section, about four-fifths of a standard deviation, with the pretest held constant. The gain lay where STOP and DARE teach: the SRSD section learned to state the other side's argument and answer it, gave more reasons with explanations, and stated its position more clearly. Its language and conventions rose no more than the comparison section's, 0.23 points against 0.24, and SRSD as taught here did not address them.

How it compares with the research. Our effect, g = 0.80, is close to the 0.82 that Graham and Perin (2007) found for writing strategy instruction with adolescents and below the 1.14 of the SRSD studies they reviewed; they also found strategy instruction more effective with struggling writers (1.02) than in regular classrooms (0.70), and ours was a regular class. Our comparison section was taught persuasive writing for the same 14 periods, so the difference is SRSD's added effect over the process writing the school already used, not over no writing instruction. Festas et al. (2015) found SRSD's gains in the elements of an essay; ours lay in the counterclaim and the reasons, the elements DARE names.

What the scoring adds. Two raters who did not know which section or occasion an essay came from scored all 170 essays, and the result held with either rater alone. A difference found by one teacher scoring her own learners would not have answered the doubt that she saw what she hoped to see. The

raters needed two calibration rounds before they agreed well enough to start, which is why training is part of the method and not a formality.

What else could explain the difference. The sections were intact, and a coin toss between two sections does not balance them as random assignment of learners would, although their pretests differed by only 0.04 standard deviations. The same teacher taught both and knew which section used SRSD, so her enthusiasm may have differed; the observations found the comparison section's lessons free of SRSD elements but cannot measure enthusiasm. With one section in each condition, the strategy cannot be separated from anything else that differed between the two sections that quarter, such as the hour they met. Learners of the two sections talk to each other; if the comparison section heard of the strategy, the difference would be smaller, not larger. The SRSD section's essays were also longer, and raters can reward length; the rubric scored the parts of an argument and their quality, not the number of words, but this cannot be ruled out.

Limits. The posttest came five days after the last lesson, so the study shows what SRSD did at the end of the unit, not whether it lasted. The Counterclaim trait's pretest scores sat mostly at 1, which kept its pretest kappa low; a level between "No mention of the other side" and "Mentions that others disagree" would spread those scores. Exact agreement on the study's essays, 59.6%, was lower than the 73.3% the raters reached on 12 practice essays in their second calibration round; a calibration check midway through the scoring would show such a drop while it can still be corrected.

What we would change. We would add an editing step with a short conventions checklist to the SRSD lessons, give a delayed essay in the following quarter, and, with more sections, assign several sections to each condition so that the analysis can treat the section as the unit.

Conclusion. Self-Regulated Strategy Development, taught in 14 writing periods with its six stages and its self-regulation procedures, improved the persuasive essays of Grade 8 learners on the parts of an argument it teaches. The comparison section is receiving it in Quarter 3, and the department will use the rubric and the anchor essays in its own scoring.

Action Plan

Table 7

Action Plan

Objective	Activities	Persons involved	Timeline	Resources (PhP)	Success indicator
Give the comparison section SRSD	SRSD with STOP and DARE in the Quarter 3 opinion editorial lessons, with the same stages and records	First researcher	November 4, 2025, to January 23, 2026	640 (cue cards, planning sheets)	14 lessons held; essays scored blind with the same rubric
Teach conventions within SRSD	Add an editing step with a five-item conventions checklist to the revising stage, in both sections	First researcher, English department head	Quarter 3, SY 2025–2026	0	Language and Conventions trait rises on the Quarter 3 essay
Find out whether the gain lasts	A delayed persuasive essay in Quarter 4 on a new prompt, scored blind by the same raters	Researchers, raters	March 2026	300 (printing)	SRSD section's mean stays above its posttest level
Make blind double scoring a department practice	LAC session on scoring with the rubric and anchors; each English teacher scores 12 practice essays and compares	English department, school head	February 2026	420 (printing)	Teachers reach 70% exact agreement in the second round
Test SRSD across more sections	Proposal for a study of four Grade 8 sections, two drawn by lot for SRSD, in SY 2026–2027	Researchers, school research committee	Next call for proposals	Up to 30,000 (BERF)	Proposal submitted and approved

References

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220. https://doi.org/10.1037/h0026256

De La Paz, S., & Graham, S. (1997). Effects of dictation and advanced planning instruction on the composing of students with writing and learning problems. *Journal of Educational Psychology, 89*(2), 203–222. https://doi.org/10.1037/0022-0663.89.2.203

Department of Education. (2012). *DepEd child protection policy* (DepEd Order No. 40, s. 2012). https://www.deped.gov.ph/wp-content/uploads/2012/05/DO_s2012_40.pdf

Department of Education. (2016). *Adoption of the Basic Education Research Agenda* (DepEd Order No. 39, s. 2016). https://www.deped.gov.ph/wp-content/uploads/2016/06/DO_s2016_039.pdf

Department of Education. (2017). *Research management guidelines* (DepEd Order No. 16, s. 2017). https://www.deped.gov.ph/wp-content/uploads/2017/03/DO_s2017_016.pdf

Department of Education. (2023). *MATATAG K to 10 curriculum of the K to 12 program: English, Grades 2–10*. https://www.deped.gov.ph/wp-content/uploads/2-FINAL-MATATAG-English-CG-2023-Grades-2-10-withAppendices.pdf

Department of Education. (2024a). *Amendment to DepEd Order No. 010, s. 2024* (DepEd Order No. 012, s. 2024). https://www.deped.gov.ph/wp-content/uploads/DO_s2024_012.pdf

Department of Education. (2024b). *Policy guidelines on the implementation of the MATATAG curriculum* (DepEd Order No. 010, s. 2024). https://www.deped.gov.ph/wp-content/uploads/DO_s2024_010.pdf

Department of Education. (2025). *Implementing guidelines for the Academic Recovery and Accessible Learning Program in Reading for Key Stages 1 to 3* (DepEd Memorandum No. 064, s. 2025). https://www.deped.gov.ph/wp-content/uploads/DM_s2025_064r.pdf

Department of Education. (2026). *Foundational guidelines on artificial intelligence in basic education* (DepEd Order No. 003, s. 2026). https://www.deped.gov.ph/2026/02/20/february-20-2026-do-003-s-2026-foundational-guidelines-on-artificial-intelligence-ai-in-basic-education/

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43*(6), 543–549. https://doi.org/10.1016/0895-4356(90)90158-L

Festas, I., Oliveira, A. L., Rebelo, J. A., Damião, M. H., Harris, K., & Graham, S. (2015). Professional development in self-regulated strategy

development: Effects on the writing performance of eighth grade Portuguese students. *Contemporary Educational Psychology, 40*, 17–27. https://doi.org/10.1016/j.cedpsych.2014.05.004

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*(3), 613–619. https://doi.org/10.1177/001316447303300309

Graham, S., Harris, K., & Hebert, M. A. (2011). *Informing writing: The benefits of formative assessment* [A Carnegie Corporation Time to Act report]. Alliance for Excellent Education. https://files.eric.ed.gov/fulltext/ED537566.pdf

Graham, S., McKeown, D., Kiuhara, S., & Harris, K. R. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology, 104*(4), 879–896. https://doi.org/10.1037/a0029185

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99*(3), 445–476. https://doi.org/10.1037/0022-0663.99.3.445

Harris, K. R., Graham, S., MacArthur, C. A., Reid, R., & Mason, L. H. (2011). Self-regulated learning processes and children's writing. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 187–202). Routledge. https://doi.org/10.4324/9780203839010.ch12

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70. https://www.jstor.org/stable/4615733

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Republic Act No. 10173. (2012). *Data Privacy Act of 2012*. https://lawphil.net/statutes/repacts/ra2012/ra_10173_2012.html

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 92–96). https://doi.org/10.25080/Majora-92bf1922-011

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

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research, and Evaluation, 9*, Article 4. https://doi.org/10.7275/96jp-xz07

UNICEF & SEAMEO. (2020). *SEA-PLM 2019 main regional report: Children's learning in 6 Southeast Asian countries*. https://www.seaplm.org/PUBLICATIONS/regional%20results/SEA-PLM%202019%20Main%20Regional%20Report.pdf

What Works Clearinghouse. (2022). *What Works Clearinghouse procedures and standards handbook, version 5.0* (WWC 2022008). U.S. Department of Education, Institute of Education Sciences. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/Final_WWC-HandbookVer5.0-0-508.pdf

Financial Report

Table 8

Statement of Expenses

Date	Particulars	Receipt no.	Amount (PhP)	Category
August 20, 2025	Consent letters in English and Filipino and assent scripts, 95 copies	OR 3317	190	Reproduction
August 26, 2025	Prompts, planning sheets, and answer sheets for the pretest, 100 sets	OR 3352	300	Reproduction
September 1, 2025	STOP and DARE cue cards and planning sheets for 46 learners	OR 3389	640	Supplies and materials
October 20, 2025	Prompts, planning sheets, and answer sheets for the posttest, 100 sets	OR 3561	300	Reproduction
November 6, 2025	Rubrics, anchor essays, and 510 scoring sheets for three raters	OR 3604	420	Reproduction
November 7, 2025	Snacks for the rater training session, 6 persons	OR 1180	600	Food
January 28, 2026	Printing and binding of the report, 3 copies	OR 3722	900	Reproduction
Total			**3,350**	

Note. Fund source: the researchers' own funds. No school funds or grants were used. Receipt numbers are invented for this sample.

Appendices

Appendix A. Data Set and Notebook

The Excel data set (the learners, both raters' scores of all 170 essays, the third rater's scores, and every statistic Excel can compute as a formula) and the Jupyter notebook that reproduces every result.

Appendix B. Consent Letter and Assent Script

Kept with the school's copy, with the signed forms, and left out of the archived report.

Appendix C. Analytic Rubric and Scoring Sheet

The five traits, each with its four descriptors, and the scoring sheet the raters used for each essay.

Table 9

Analytic Rubric for the Persuasive Essay

Trait	4	3	2	1
Claim (EN8INF-II-16)	States a clear position on the question in the opening and keeps it to the end; the ending restates it.	States a clear position, but late, or the ending does not return to it.	The position can be worked out but is never stated plainly, or it shifts.	Takes no position on the question, or takes both sides.
Reasons and evidence (EN8INF-II-17, EN8INF-II-6)	Three or more reasons, each explained with a fact, an example or an experience that supports it.	Two or more reasons, at least two of them explained.	Reasons are listed, but only one is explained, or the explanations only repeat the reason.	No reason, or reasons that do not support the position.
Counterclaim (EN8INF-II-17, EN8INF-II-6)	States an argument of the other side fairly and answers it with a reason or evidence.	States an argument of the other side and answers it, but the answer only contradicts it.	Mentions that others disagree but does not say what they argue, or states it without an answer.	No mention of the other side.

Trait	4	3	2	1
Organization (EN8INF-II-2, EN8INF-II-19)	An opening that introduces the issue, body ideas in a logical order joined by transitions, and an ending that closes the argument.	Opening, body and ending are present; transitions are few or repetitive.	Some order, but one of the three parts is missing or the ideas jump.	No clear order; a list of sentences.
Language and conventions (EN8INF-II-20, EN8INF-II-21)	Varied sentences and precise words; few errors, none that blur the meaning.	Mostly correct sentences and words; errors do not blur the meaning.	Frequent errors in grammar, spelling or punctuation; the meaning is sometimes blurred.	Errors in most sentences; the meaning is often hard to follow.

Note. Each trait is scored 1 to 4; the total is 5 to 20. Competency codes are from the MATATAG English curriculum, Grade 8, Quarter 2.

Appendix D. The Two Prompts

Prompts A and B, with their instructions, as given to the learners.

Prompt A. Your school head is deciding whether every Grade 8 learner should be required to join at least one school club. Write an essay to persuade the school head. State your position, give your reasons, and answer at least one argument of people who disagree with you. You have 40 minutes.

Prompt B. Your school head is deciding whether every Grade 8 learner should be required to take care of a part of the school's vegetable garden for one quarter. Write an essay to persuade the school head. State your position, give your reasons, and answer at least one argument of people who disagree with you. You have 40 minutes.

Appendix E. Lesson Outline and Fidelity Checklists

The 14 SRSD lessons by stage, the components the observer checked in each, and the comparison section's lesson outline.

Appendix F. Declaration of AI Use

We, the researchers, declare that we used [name of the AI tool] only to check the grammar of the English text of this report, for the purpose of finalizing the research study. The prompt used was: "Check the grammar of this paragraph. Do not change its meaning or add content." We have reviewed, edited, and understood all the content, and we take full accountability for it. No essays, names, scores, or other personal information were entered into the AI tool (adapted from Annex A of DepEd Order No. 003, s. 2026).