Reciprocal Teaching and Reading Comprehension of Grade 7 Learners in English

An Action Research

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

Conducted June to September 2025

Abstract

This action research tested whether reciprocal teaching, in which learners take turns leading a small group through predicting, questioning, clarifying, and summarizing, would improve the reading comprehension of Grade 7 learners in English. One teacher's four sections in a public secondary school, all 162 learners, took part in School Year 2025–2026. In a stepped-wedge design analyzed as a multiple baseline across sections, the strategy started in Sections A, B, C, and D in weeks 3, 5, 7, and 9, the order drawn by lot, and every learner took a 10-item comprehension probe each week for 12 weeks. A linear mixed-effects model with a time trend, section effects, the intervention, and a random intercept and slope for each learner estimated the effect, and a randomization test over the 24 possible start orders checked it. The learners took 1,872 of 1,936 probes. Reciprocal teaching raised scores by 0.82 points of 10, 95% CI [0.57, 1.08], d = 0.32; the real order gave the largest of the 24 estimates, one-sided p = .042, the smallest four sections allow. Each section rose 0.88 to 1.08 points over its first two weeks of the strategy, while the sections still waiting changed by −0.09 to 0.61. The effect grew over two weeks and was largest on inferential items. Learners' weekly growth varied, but one learner's 12-week slope had a reliability of only .34. Reciprocal teaching is recommended for all Grade 7 English sections, with a delayed probe and a two-teacher study to follow.

Keywords: reciprocal teaching, reading comprehension, multiple-baseline design, stepped-wedge design, mixed-effects model

Acknowledgment

The researchers thank the school head, who approved the study and kept the four sections' English periods unchanged for the quarter; the three English teachers who reviewed every probe and its answer key before it was used; the Grade 7 advisers, who gave the consent forms out and collected them; the parents, who gave their consent at the Grade 7 parents' orientation; and the 162 learners, who read a new passage every week for 12 weeks. The study was funded by the researchers.

Reciprocal Teaching and Reading Comprehension of Grade 7 Learners in English

Context and Rationale

Filipino learners read far below what their age calls for. In the Programme for International Student Assessment (PISA) of 2022, the Philippines averaged 347 points in reading, and only 24% of its 15-year-olds reached Level 2, the baseline of proficiency, against 74% across the Organisation for Economic Co-operation and Development (OECD, 2023). The 2025 round showed a rise, to 367 points and 31% at Level 2 or higher, still less than half the OECD average of 69% (OECD, 2026). Reading comprehension in English decides how much a learner gets from every subject taught in it.

Our four Grade 7 sections showed the same need. On the beginning-of-year Philippine Informal Reading Inventory (Phil-IRI), which DepEd now gives in Grades 4 to 10 (Department of Education [DepEd], 2025a), 16 of our 162 learners read English at the Frustration level. In class, many learners could answer questions whose answers stood in the text but not questions that asked them to infer, and few checked their own understanding as they read. The MATATAG curriculum asks Grade 7 learners to "analyze literary texts as expressions of individual or communal values" from the first quarter (DepEd, 2023, EN7LIT-I-1), which assumes they can first understand what they read.

Reciprocal teaching was designed for such learners. Palincsar and Brown (1984) taught seventh-grade poor comprehenders to take turns leading a dialogue on a text with four activities, "summarizing (self-review),

questioning, clarifying, and predicting," and found "sizable gains on criterion tests of comprehension," maintenance over time, and generalization to classroom comprehension tests. A review of 16 studies found a median effect size of 0.32 on standardized tests and 0.88 on comprehension tests the experimenters developed (Rosenshine & Meister, 1994). In whole classrooms, effects are smaller: a meta-analysis of 52 studies of reading-strategy instruction taught to whole classes found an effect of d = 0.186 on standardized tests and 0.431 on researcher-developed tests, with larger effects in Grades 6 to 8 and where researchers rather than teachers gave the instruction, and smaller ones in studies that did not model the nesting of learners in classes (Okkinga, van Steensel, van Gelderen, van Schooten, et al., 2018). A year-long trial with low-achieving adolescents found that reciprocal teaching helped "only when experimental teachers provided high-quality strategy instruction" (Okkinga, van Steensel, van Gelderen, & Sleegers, 2018). Philippine evidence is thin: a quasi-experiment with 20 Grade 10 learners found no significant effect (Rivera & Taglucop, 2019), and a study of 120 Grade 9 learners reported a significant gain (Sevilla et al., 2025).

Whether reciprocal teaching helps Grade 7 learners in a public school's regular English periods, taught by their own teacher with the quarter's texts, is therefore a question for our own classrooms. A teacher with four sections also faces a practical problem: holding one section back for a whole quarter as a comparison group denies it a strategy the teacher expects to help. A staggered start answers both: every section gets the strategy within the quarter, and the sections still waiting show what the weeks alone would have

done (Hemming et al., 2015). The study addresses the Teaching and Learning theme of the Basic Education Research Agenda (DepEd, 2016).

Innovation, Intervention, and Strategy

The innovation was reciprocal teaching, brought into the regular English periods of four Grade 7 sections one section at a time. Everything else stayed the same: the teacher, the texts and competencies of each week, the time of each period, and the weekly probe. In the first quarter, the lessons' texts were Philippine poems; from August 26, the second quarter's short stories (DepEd, 2023). Reciprocal teaching was used on those texts and on short prose pieces that went with them, such as a poet's life or the background of a poem.

The Four Strategies

Learners read a text in short parts and, after each part, took turns leading their group through four strategies: predicting what the next part would say, asking a question about the part just read, clarifying a word or idea that was unclear, and summarizing the part in a sentence or two (Palincsar & Brown, 1984). Each strategy had a role card with sentence starters in English, such as "I think the next part will tell us…", "Why did…?", "I didn't understand… so I…", and "This part was mostly about…".

The Gradual Release

In a section's first week of the strategy, the teacher modeled each strategy aloud on the week's text, one strategy a period, then all four together. In the second week, the learners worked in groups of four with role cards while the teacher led one group at a time and prompted the others. From the

third week, the groups ran on their own, the roles rotating each session, while the teacher listened, corrected, and modeled again where a group was stuck. Reciprocal teaching took about 25 minutes of three English periods a week; the rest of each period followed the quarter's lessons.

Before a Section's Turn

Until its start week, a section was taught the same texts and competencies in the usual way: the teacher introduced the text, the learners read it, and the teacher asked comprehension questions to the class. No section was told when its turn would come.

Strategy and Evaluation Plan

The design started the strategy in Section A in week 3, B in week 5, C in week 7, and D in week 9, the order drawn by lot before the study, so that every section received it and each start could be compared with the sections still waiting. A 10-item probe each week measured comprehension in all four sections. The department head observed one reciprocal teaching session a week in each section using the strategy, with a 10-component checklist. Three criteria for success were set in advance: at least eight of the ten components seen in the observed sessions on average; a strategy effect significant at the .05 level in the mixed model, with a randomization test that agreed; and a rise in each section after its own start that was larger than the change in the sections still waiting.

Action Research Questions

Main question: Does reciprocal teaching improve the reading comprehension of Grade 7 learners in English?

1. What were the four sections' scores on the weekly comprehension probe before and during reciprocal teaching?
2. Is there a significant difference between the probe scores with and without reciprocal teaching, allowing for the weeks' trend, the sections' levels, and each learner's own starting point and growth?
3. Did each section's scores rise when its own turn came, and not before?
4. How did the effect change over a section's first weeks of the strategy, and on which kinds of questions?
5. How much did the learners differ in their weekly growth?

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

$H_0$1: There is no significant difference between the learners' probe scores with and without reciprocal teaching.

$H_0$2: Reciprocal teaching had no effect on any learner's probe score, so the real order of the sections' starts is no more likely than any of the other 23 orders to give the largest estimated effect.

Action Research Methods

Research Design

The study used a stepped-wedge design, analyzed as a multiple-baseline design across sections. In a stepped-wedge design, every group receives the intervention, one group at a time, in an order set in advance, and measurements continue through every step (Hussey & Hughes, 2007; Hemming et al., 2015). In a multiple-baseline design across groups, the intervention starts in one group while the others are still in baseline, so each start can show its effect at a different point in time, and the What Works

Clearinghouse counts a design with at least three such starts "at three different points in time" as able to show an effect (What Works Clearinghouse [WWC], 2022, p. 111). Four sections and four start dates gave four such demonstrations. The design suited a teacher who wanted every section to receive the strategy within the quarter and could not hold one back as a comparison group.

The order in which the sections started was drawn by lot from the four fixed start weeks before the study. This made a randomization test possible: if reciprocal teaching had no effect, any of the 24 orders the lot could have drawn would be as likely as the real one to show the largest effect, the logic of the randomization tests developed for multiple-baseline designs (Koehler & Levin, 1998; Kratochwill & Levin, 2010).

Participants and Other Sources of Data and Information

The participants were the 162 learners of the teacher's four Grade 7 sections in a public secondary school in School Year 2025–2026: 41 in Section A, 39 in B, 42 in C, and 40 in D; 71 girls and 91 boys. All learners of the four sections took part; no sampling was done. 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, 16 learners read English at the Frustration level; they began ARAL-Reading sessions outside their English periods when the program started in the second quarter, on August 26, 2025, in study weeks 11 and 12 (DepEd, 2025a). Two learners, one in Section B and one in Section D, transferred out after week 8; their probes up to that week were kept.

Other sources of data were the weekly probes' item responses, the department head's fidelity checklists, and the school's records of the learners in ARAL-Reading.

Data Gathering Methods

The Weekly Comprehension Probe

Each week, every learner read a new passage of 300 to 350 words at Grade 7 level, six of them literary and six informational, written or adapted by the teacher and never used in the lessons, and answered 10 multiple-choice items: four literal, four inferential, and two critical. Passages the strategy had not been practiced on test whether the learners could use it on their own, and keep the probe from favoring the sections that had already read a lesson's text (WWC, 2022, p. 27). The teacher wrote the 12 probes before the study; three English teachers reviewed every item for its match to the passage and to its kind, and the teacher revised the items they flagged. All four sections took the same probe in the same week, in the Thursday English period, in 15 minutes. The probes did not count toward the learners' grades. Papers were collected and discussed only after all four sections had taken them, so that no section saw the answers early.

The probe was given on Thursdays, from June 19 to September 4, 2025, except in week 10, when it was given on Tuesday, August 19, because the Quarter 1 examination fell on August 20 and 22 and August 21 was a holiday (DepEd, 2025b). Classes were held in all 12 weeks: the school's province was not among those where classes were suspended for the monsoon rains of July 21 to 25, 2025.

Reliability

Each week's internal consistency was computed with the Kuder-Richardson formula 20 (KR-20) on the items as answered. The values ranged from .66 to .75, with a median of .70, above the .60 the What Works Clearinghouse sets for internal consistency (WWC, 2022, p. 27): adequate for comparing groups and weeks, too low for judging one learner on one probe, which is why the analysis pooled each learner's 12 probes.

Fidelity

The department head, who co-led the study, observed one reciprocal teaching session a week in each section that was using the strategy, except in the examination week, and checked ten components: the four strategies, the role cards, the teacher's modeling, think-alouds, the gradual release, groups of four, and the groups' shared summary.

Recording and Storage

Answer sheets were scored by the teacher with the key and entered into an Excel file under learner codes on the day of the probe. The key that links codes to names was kept apart from the data, in the school's office, and the sheets were stored in a locked cabinet.

Ethical Considerations

The school head approved the study in writing in May 2025, the parents gave written consent at the Grade 7 parents' orientation on June 13, 2025, before classes opened, and each learner gave assent in the first English period. The study followed the Research Management Guidelines (DepEd, 2017) and the Child Protection Policy (DepEd, 2012). The probes did not count

toward grades, and every section received the strategy within the quarter. Learners were identified only by codes. The researchers' use of an artificial intelligence (AI) tool is declared in Appendix D (DepEd, 2026).

Data Analysis

For the first question, the probe scores were described by section, week, and phase. For the second, the main analysis was a linear mixed-effects model, fitted by restricted maximum likelihood in the statsmodels library for Python (Seabold & Perktold, 2010): the probe score as the outcome; the week (centered at its midpoint) as a time trend; the section; and the strategy, coded 1 in a section's weeks from its start; with a random intercept and a random slope on the week for each learner. The strategy's coefficient is its effect in points, holding constant the weeks' trend, the sections' levels, and each learner's own starting point and growth. The effect size d is that coefficient over the pooled within-week standard deviation of the scores before the strategy. Three checks followed: a model with a separate effect for each week in place of the linear trend, which also absorbs the passages' differences in difficulty; the main model without the learners in ARAL-Reading; and a regression on the 48 section-week means with section and week effects, which treats each section-week as one observation.

Because a model fitted to learners' scores counts 162 learners but the strategy was given to only four sections, its standard errors may understate the uncertainty (Cameron & Miller, 2015; Taljaard et al., 2016). The randomization test does not depend on that assumption. The model with week effects was refitted for each of the 24 orders the lot could have drawn, with

the strategy switched on as that order would have it; the one-sided *p* value is the share of the 24 estimates at least as large as the real order's. With four sections it can be no smaller than 1/24, or .042.

For the third question, each section's change from the two weeks before its start to its first two weeks with the strategy was set beside the change in the sections still waiting over the same weeks. For the fourth, the week-effects model was refitted with the strategy split into a section's first week, second week, and third week onward, and the main model was fitted to each subscore. For the fifth, each learner's growth slope was the model's weekly trend plus the learner's own random slope; how far one learner's slope can be trusted was estimated as the slopes' variance over itself plus the sampling variance of a 12-week slope.

All analyses were run in Python 3.12 with pandas and statsmodels, free software a school can use without a license; the notebook and the Excel data set reproduce every result (Appendix A).

Discussion of Results and Reflection

Implementation

The learners took 1,872 of the 1,936 probes they could have taken (96.7%); most missing probes were single absences. In the 24 sessions the department head observed, a mean of 8.5 of the 10 components was seen (84.6%): 88.9% in Section A, 82.9% in B, 84.0% in C, and 76.7% in D. In each section's first week of the strategy, the mean was 72.5%; the groups needed a week or two to run on their own. Section D was observed only in its first three weeks with the strategy, when every section scored lower; over their first three

observations, Sections A, B, and C averaged 80.0%, 76.7%, and 76.7%. The first criterion for success, eight of ten components on average, was met overall and in three of the four sections.

Question 1: Probe Scores Before and During the Strategy

Before the strategy, the four sections averaged 4.6 to 5.3 points of 10; with it, 5.4 to 6.7 (Table 1 and Figure 1). Every section's mean rose, by 0.9 to 1.4 points.

Table 1

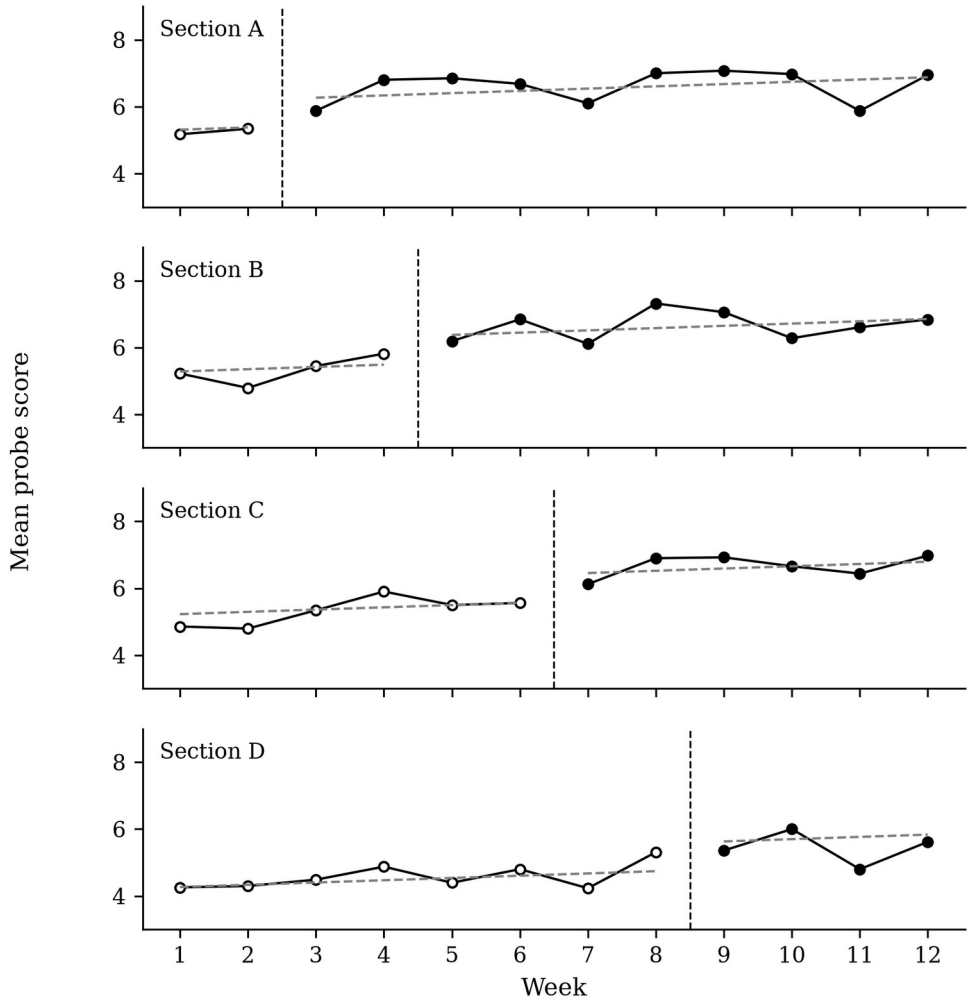
Probe Scores Before and During Reciprocal Teaching, by Section

Section	Start week	Learners	Before: *M* (*SD*)	Before: probes	During: *M* (*SD*)	During: probes	Difference
A	3	41	5.26 (2.26)	81	6.62 (2.31)	401	1.36
B	5	39	5.32 (2.42)	148	6.66 (2.22)	297	1.34
C	7	42	5.33 (2.45)	246	6.67 (2.17)	245	1.34
D	9	40	4.59 (2.73)	309	5.45 (2.85)	145	0.86
All		162	5.03 (2.56)	784	6.48 (2.37)	1,088	1.46

Note. Scores are of 10. The differences are raw: they mix the strategy's effect with the weeks' trend and the passages' differences in difficulty, which the models separate.

Figure 1

Mean Probe Score by Week in the Four Sections



Note. Open markers show weeks before the section used reciprocal teaching; filled markers, weeks with it. Vertical dashed lines mark each start. Dashed lines without markers show the mixed model's fitted means.

Question 2: The Strategy's Effect

The mixed model estimated that reciprocal teaching raised probe scores by 0.82 points of 10, 95% CI [0.57, 1.08], $z = 6.28$, $p < .001$, $d = 0.32$, 95% CI [0.22, 0.42], holding constant a weekly trend of 0.068 points and the sections' levels (Table 2). $H_0 1$ is rejected. Section D started 1.05 points below Section

A; Sections B and C did not differ from A. Learners differed more from one another than from week to week: the intraclass correlation was .58.

Table 2

Linear Mixed-Effects Model of the Weekly Probe Score

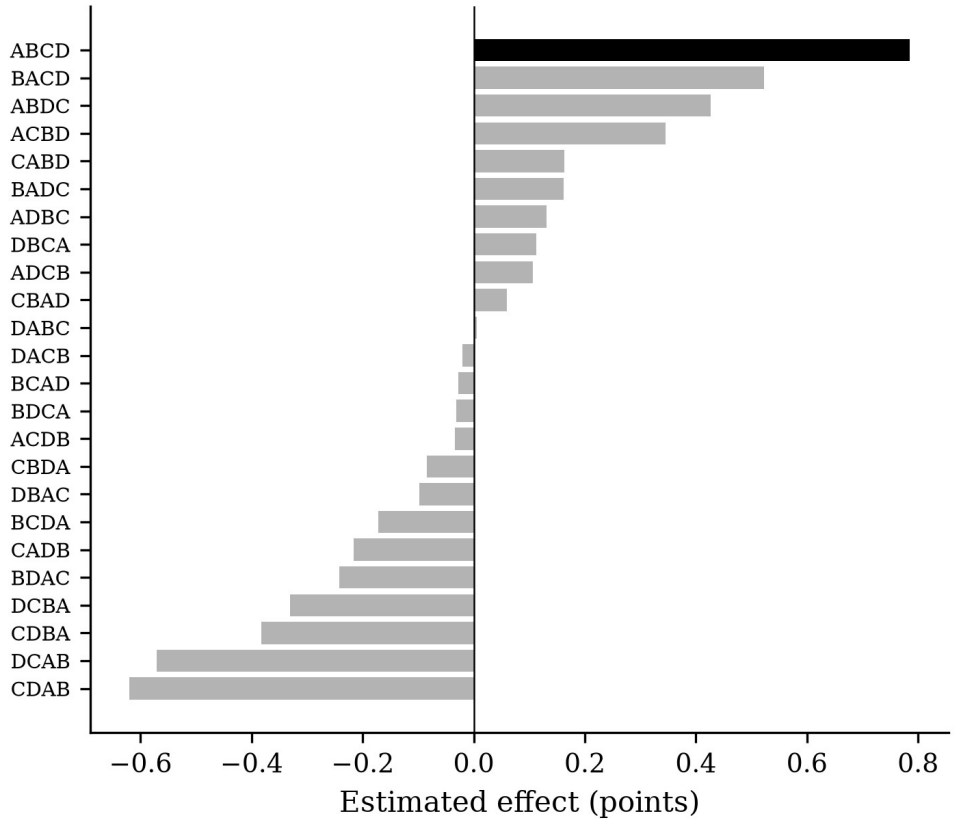
Term	*b*	*SE*	*z*	*p*	95% CI
Intercept (Section A, mid-study, without the strategy)	5.68	0.32	17.93	< .001	[5.06, 6.31]
Reciprocal teaching	0.82	0.13	6.28	< .001	[0.57, 1.08]
Week (each week)	0.068	0.019	3.62	< .001	[0.031, 0.105]
Section B (vs. A)	−0.03	0.43	−0.07	.946	[−0.87, 0.81]
Section C (vs. A)	−0.09	0.42	−0.20	.840	[−0.91, 0.74]
Section D (vs. A)	−1.05	0.43	−2.44	.015	[−1.89, −0.21]

Note. 1,872 probes of 162 learners. Random effects: learners' intercepts, variance 3.42; learners' weekly slopes, variance 0.0088 (*SD* 0.094); their correlation, −.01; residual variance 2.43. Restricted maximum likelihood; the week is centered at 6.5.

The checks agreed. With a separate effect for each week, the estimate was 0.79, 95% CI [0.52, 1.05]; without the 16 learners in ARAL-Reading, 0.85, 95% CI [0.58, 1.12]; and on the 48 section-week means, 0.80 (*SE* 0.14). In the randomization test, the real order gave the largest estimate of the 24 orders, 0.79; the next was 0.52 (Figure 2). The one-sided p is 1/24 = .042, the smallest four sections allow, and $H_0 2$ is rejected. The mixed model's $p < .001$ rests on treating learners as independent beyond their own random effects; the randomization test's .042 does not, and it is the figure to lean on.

Figure 2

Estimated Effect Under Each of the 24 Orders in Which the Sections Could Have Started



Note. Each bar is the strategy's coefficient in the model with week and section effects, refitted with the strategy switched on as the order would have it. The black bar is the real order, A, B, C, D.

Question 3: Each Start as a Test

Each section's mean rose over its first two weeks of the strategy by 0.88 to 1.08 points (Table 3). Over the same weeks, the sections still waiting changed by −0.09 to 0.61. Section A's start is the weakest test of the first three: its baseline had only two probes, and the waiting sections also rose by 0.61 over weeks 3 and 4, a sign that part of A's rise came with the weeks, which the models' time trend and week effects take out. Section D's start had

no section left waiting; its comparison rests on the sections already using the strategy, which changed by 0.23, and on the time trend.

Table 3

Change Over Each Section's Start, Beside the Other Sections

Start	Section's weeks compared	Section's change	Sections still waiting	Their change	Sections already using it	Their change
Week 3, A	1–2 vs. 3–4	1.08	B, C, D	0.61	None	
Week 5, B	3–4 vs. 5–6	0.88	C, D	−0.09	A	0.42
Week 7, C	5–6 vs. 7–8	0.98	D	0.17	A, B	0.00
Week 9, D	7–8 vs. 9–10	0.90	None		A, B, C	0.23

Note. Changes are differences in mean score, of 10, between the two weeks from the start and the two weeks before it (for Section A, weeks 1 and 2, its whole baseline). The third criterion for success is met at the first three starts (at the first by the narrowest margin, since the waiting sections also rose 0.61) and cannot be judged at the fourth.

Question 4: How the Effect Grew, and on Which Questions

The effect grew over a section's first two weeks of the strategy and then held (Table 4). In the first week, while the teacher modeled the strategies, scores were 0.56 points higher than the week and section alone would predict; in the second week, 0.98; from the third week on, 0.93. Of the three kinds of questions, inferential items gained the most for each item: 0.40 points over four items (0.10 an item), against 0.30 over four literal items (0.08 an item) and 0.13 over two critical items (0.06 an item).

Table 4

The Strategy's Effect by Week of the Strategy and by Kind of Question

Estimate	*b*	95% CI	*p*
First week of the strategy (158 probes)	0.56	[0.23, 0.89]	< .001
Second week (156 probes)	0.98	[0.65, 1.31]	< .001
Third week and later (774 probes)	0.93	[0.57, 1.28]	< .001
Literal items (4 items)	0.30	[0.16, 0.45]	< .001
Inferential items (4 items)	0.40	[0.24, 0.55]	< .001

Estimate	*b*	95% CI	*p*
Critical items (2 items)	0.13	[0.02, 0.24]	.021

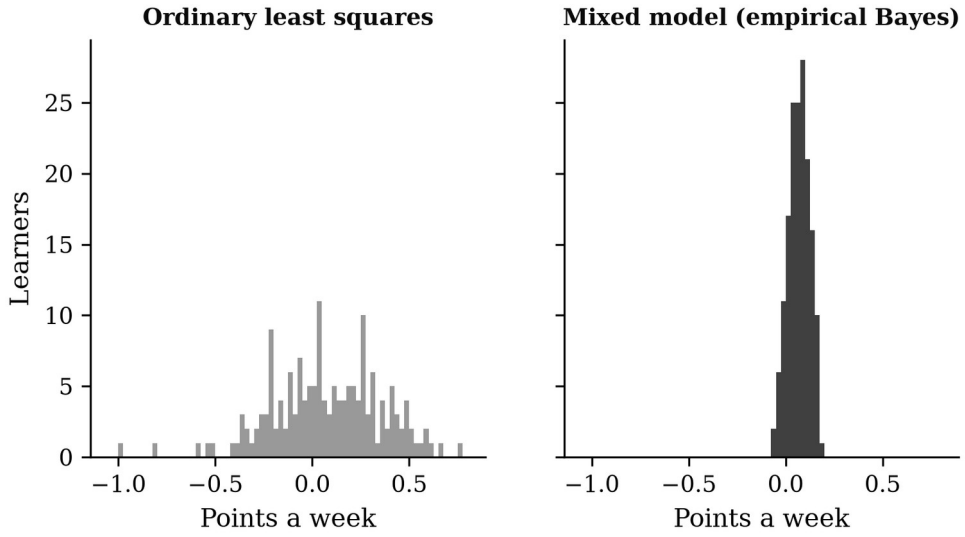
Note. The first three rows come from the model with week and section effects, the strategy split by its week; the last three from the main model fitted to each subscore.

Question 5: The Learners' Growth

The average learner gained 0.068 points a week apart from the strategy, or about 0.75 points over the 12 weeks. The model estimated that learners' weekly growth varied with a standard deviation of 0.094 points. Each learner's own slope from the model ranged from −0.05 to 0.19 points a week, and 143 of the 162 were positive (Figure 3). A learner's own straight line through 12 probes was much noisier, from −1.0 to 0.76 points a week: with a probe of 10 items, the reliability of one learner's 12-week slope was only .34. The study can describe how much learners' growth differs, but it cannot rank one learner's growth against another's.

Figure 3

Learners' Weekly Growth: Their Own Lines and the Mixed Model's Slopes



Note. The left panel shows each learner's least-squares line through their probes, with the strategy as a second term; the right, the model's empirical Bayes slopes, drawn toward the average where a learner's own line is noisy.

Reflection

What we found. Reciprocal teaching raised the weekly probe scores of all four sections, by 0.82 points of 10 when the weeks' trend and the sections' levels were held constant. Each section rose when its own turn came, and the real start order gave the largest effect of the 24 the lot could have drawn. The effect took two weeks to build, while the learners learned the roles, and was largest on inferential questions, the kind the strategies of questioning and clarifying practice.

How it compares with the research. An effect of $d = 0.32$ on our own probes is smaller than the median of 0.88 that Rosenshine and Meister (1994) found on experimenter-developed tests and lies between the 0.186 and 0.431 that Okkinga, van Steensel, van Gelderen, van Schooten, et al. (2018) found for strategy instruction in whole classes on standardized and researcher-developed tests. Their review found smaller effects where teachers, not researchers, gave the instruction, as in our study. Our probes used passages the strategy had not been practiced on, which also keeps the effect honest.

What else could explain the gains. The weeks themselves raised scores: every section rose from weeks 1 and 2 to weeks 3 and 4, and three of the four dipped in week 11, the first week of the second quarter, Section B a week earlier. The models hold these changes constant, but a section-specific event in the week of its start, such as a section's own activity or a substitute teacher, could not be separated from the strategy. The learners in the sections that had started may also have talked about the strategy with those still waiting, which would make the comparison smaller, not larger. ARAL-Reading

sessions for the 16 learners at the Frustration level began only in the second quarter, in weeks 11 and 12, when all four sections were already using the strategy; leaving those learners out changed the estimate little, from 0.82 to 0.85.

Limits. Four sections allow a randomization *p* no smaller than .042, and a mixed model fitted to 162 learners in four sections can understate its uncertainty (Cameron & Miller, 2015; Taljaard et al., 2016). Section A's baseline had only two probes, but Sections B, C, and D gave three baselines of four or more probes, enough for the What Works Clearinghouse's (2022) multiple-baseline standards with reservations. No schedule of 12 weekly probes gives three start weeks with six probes before each and five after, so the study cannot meet those standards without reservations. Section D used the strategy for four weeks, through the examination and the turn of the quarter, and with the lowest fidelity. The probes, at a KR-20 of about .70, were reliable enough for sections and weeks but not for judging one learner's growth. The same teacher taught all four sections, so the study shows what reciprocal teaching did in her hands.

What we would change. We would give every section at least three baseline probes, draw the start weeks themselves by lot and not only the order, add items to each probe, and keep a probe going for some weeks after the last section starts.

Conclusion. Reciprocal teaching, taught with a gradual release in the regular English periods, raised Grade 7 learners' reading comprehension on new passages by about a third of a standard deviation, most on inferential

questions, and each section's gain came when its own turn came. The strategy will continue in all four sections, and a study with a second teacher's sections would show whether the effect holds beyond one classroom.

Action Plan

Table 5

Action Plan

Objective	Audience	Activity and channel	Persons responsible	Timeline	Budget (PhP)	Evidence of success
Keep reciprocal teaching in all four sections through Quarter 2	Grade 7 learners	Three sessions a week, roles rotating; fidelity checklist every two weeks	English teacher, department head	September to October 2025	0	Eight of ten components seen; probe means kept
Share the study with the English department	English teachers, Grades 7 to 10	LAC session with the role cards, a modeled lesson and the four panels	Researchers	October 2025	300 (printing)	Two more teachers try the strategy
Follow the probes' effect later	Grade 7 learners	A delayed probe in Quarter 3 on new passages, all sections	English teacher	January 2026	600 (printing)	Scores stay above the baseline
Strengthen the probes	Researchers	Replace the three items with the lowest discrimination in each probe; add two items to each	English teacher, reviewers	November 2025	0	KR-20 of .75 or more
Test the strategy with another teacher's sections	School research committee	Proposal for a two-teacher study in SY 2026–2027, start order drawn by lot	Researchers, school head	Next call for proposals	Up to 30,000 (BERF)	Proposal submitted and scored

References

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources, 50*(2), 317–372. https://doi.org/10.3368/jhr.50.2.317

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. (2025a). *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. (2025b). *Multi-year implementing guidelines on the school calendar and activities* (DepEd Order No. 012, s. 2025). https://www.deped.gov.ph/wp-content/uploads/DO_s2025_012.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/

Hemming, K., Haines, T. P., Chilton, P. J., Girling, A. J., & Lilford, R. J. (2015). The stepped wedge cluster randomised trial: Rationale, design, analysis, and reporting. *BMJ, 350*, Article h391. https://doi.org/10.1136/bmj.h391

Hussey, M. A., & Hughes, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials, 28*(2), 182–191. https://doi.org/10.1016/j.cct.2006.05.007

Koehler, M. J., & Levin, J. R. (1998). Regulated randomization: A potentially sharper analytical tool for the multiple-baseline design. *Psychological Methods, 3*(2), 206–217. https://doi.org/10.1037/1082-989X.3.2.206

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods, 15*(2), 124–144. https://doi.org/10.1037/a0017736

OECD. (2023). *PISA 2022 results (Volume I and II): Country notes: Philippines*. OECD Publishing. https://www.oecd.org/en/publications/pisa-2022-results-volume-i-and-ii-country-notes_ed6fbcc5-en/philippines_a0882a2d-en.html

OECD. (2026). *PISA 2025 results (Volume I): Country notes: Philippines*. OECD Publishing. https://www.oecd.org/en/publications/pisa-2025-results-volume-i-country-notes_2d4ff9ea-en/philippines_3699b9d5-en.html

Okkinga, M., van Steensel, R., van Gelderen, A. J. S., & Sleegers, P. J. C. (2018). Effects of reciprocal teaching on reading comprehension of low-achieving adolescents: The importance of specific teacher skills. *Journal of Research in Reading, 41*(1), 20–41. https://doi.org/10.1111/1467-9817.12082

Okkinga, M., van Steensel, R., van Gelderen, A. J. S., van Schooten, E., Sleegers, P. J. C., & Arends, L. R. (2018). Effectiveness of reading-strategy interventions in whole classrooms: A meta-analysis.

Educational Psychology Review, 30(4), 1215–1239. https://doi.org/10.1007/s10648-018-9445-7

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1*(2), 117–175. https://doi.org/10.1207/s1532690xci0102_1

Rivera, J. M., & Taglucop, L. M. (2019). Improving reading comprehension of students through reciprocal reading: The case of Canitoan National High School. *International Journal of Sciences: Basic and Applied Research, 46*(2), 83–96. https://www.gssrr.org/JournalOfBasicAndApplied/article/view/10049

Rosenshine, B., & Meister, C. (1994). Reciprocal teaching: A review of the research. *Review of Educational Research, 64*(4), 479–530. https://doi.org/10.3102/00346543064004479

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

Sevilla, M. M., Badol, M. P., Dalumbay, J. S., & Fernandez, E. J. (2025). The use of reciprocal teaching method to reading comprehension of Grade 9 learners. *Asian Journal of Language, Literature and Culture Studies, 8*(2), 351–361. https://doi.org/10.9734/ajl2c/2025/v8i2243

Taljaard, M., Teerenstra, S., Ivers, N. M., & Fergusson, D. A. (2016). Substantial risks associated with few clusters in cluster randomized and stepped wedge designs. *Clinical Trials, 13*(4), 459–463. https://doi.org/10.1177/1740774516634316

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 6

Statement of Expenses

Date	Particulars	Receipt no.	Amount (PhP)	Category
June 10, 2025	Printing of consent and assent forms, 170 copies	OR 2051	170	Reproduction
June 13, 2025	Role cards on card stock, 44 sets of four	OR 2074	880	Supplies and materials
June 13, 2025	Printing of probes and answer sheets, weeks 1 to 6, 990 copies	OR 2074	990	Reproduction
July 25, 2025	Printing of probes and answer sheets, weeks 7 to 12, 990 copies	OR 2391	990	Reproduction
August 29, 2025	Fidelity checklists and data folders	OR 2502	260	Supplies and materials
Total			**3,290**	

Note. Fund source: the researchers' own funds. No school funds or grants were used. The receipts are kept with the school's copy of the report.

Appendices

Appendix A. Data and Notebook

The Excel data set (the probes, items, learners, fidelity, the section-week means and the start orders, with every computable statistic as a formula) and the Jupyter notebook that reproduces every number, both linked above.

Appendix B. Consent Letter and Assent Script

Kept with the school, not in the archived report.

Appendix C. Role Cards and the Fidelity Checklist

The four role cards with their sentence starters, and the ten components the department head checked.

Appendix D. Declaration of AI Use

1. 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 names, scores, or other personal or sensitive information were entered into the AI tool (adapted from Annex A of DepEd Order No. 003, s. 2026).