Check Before You Trust: Detecting Artificial Intelligence Errors in Grade 10 Science

An Action Research

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

Conducted June to August 2026

Abstract

This action research tested whether a short routine for checking explanations written by artificial intelligence (AI) would help Grade 10 learners tell correct explanations from wrong ones. In our review of the previous year's homework, 41 of the 137 answers that learners said came from a chatbot had a science error, and 33 of those were copied unchanged. In the routine, learners read an explanation an AI tool had written for the week's competency, flagged four kinds of wording that claim without showing, compared the explanation with two other sources in a triangulation grid, and decided and fixed. Two teachers' eight sections (337 learners with consent) took part in a switching-replications design: a lot within each teacher gave two sections the routine in Cycle 1 and the other two in Cycle 2, six 20-minute sessions each, in Term 1 of School Year 2026–2027. On three parallel 16-item forms, half of the explanations had one planted error, and verdicts were scored by signal detection. After Cycle 1, learners who had the routine discriminated better than those still waiting, a difference in d' of 0.69, 95% CI [0.45, 0.93], $p < .001$. The routine also made learners doubt correct explanations with flagged wording; a revised practice set removed most of this in Cycle 2, which replicated the gain. The second summative test did not differ significantly between the waves. The routine is recommended for Grade 10 Science, with practice sets in which flagged wording and errors are not paired.

Keywords: AI literacy, AI-generated explanations, signal detection, error detection, Grade 10 Science, switching replications

Acknowledgment

The researchers thank the school head, who approved the study and drew the lots; the five validators (two Master Teachers in Science, the Head Teacher of the Science department, an Education Program Supervisor in Science, and a chemistry teacher of the senior high school), who judged every draft item; the Head Teacher, who also observed 16 sessions; the Science teacher who scored a fifth of the papers a second time without knowing whose they were; the eight class advisers, who sent home and collected the consent forms and witnessed the learners' assent; and the 356 learners of Grade 10 Curie, Darwin, Faraday, Mendel, Galileo, Kepler, Newton, and Pasteur and their parents. The study was funded by the Basic Education Research Fund (school level).

Check Before You Trust: Detecting Artificial Intelligence Errors in Grade 10 Science

Context and Rationale

Our learners now ask chatbots for help with their Science homework, and DepEd allows it. Since February 2026, DepEd's guidelines on artificial intelligence have allowed learners from Grade 4 to Grade 12 to use AI tools themselves, and they ask schools to produce learners "who can critically evaluate AI-generated information, safeguard against misinformation, and participate meaningfully in an AI-driven society" (Department of Education [DepEd], 2026a, Sec. VI.2). Teachers are to embed "AI literacy and digital critical thinking skills" in the learning areas, so that learners gain "the ability to verify, question, and responsibly use AI-generated content" (Sec. VI.2.1(i)), and when they use AI for feedback, teachers "shall actively demonstrate the critical evaluation and refinement of AI-generated responses" (Sec. VI.1.2(iv)). The guidelines say what learners should be able to do; they leave how to teach it to the school. Filipino learners already use these tools often: in PISA 2025, 55% of the country's 15-year-old students reported using AI chatbots to help them learn at least once a week (OECD average: 46%), and only 5% never or almost never used them for any of the schoolwork tasks PISA asked about (Organisation for Economic Co-operation and Development [OECD], 2026, p. 9).

Our own records showed why it matters. From January 12 to 30, 2026, in Quarter 3 of School Year 2025–2026, the two of us reviewed 412 written homework answers on evolution and homeostasis from our Grade 10 classes.

A line on each homework sheet asked learners to say whether they had used a chatbot, and 137 answers (33.3%) said they had. Of these, 41 (29.9%) held a science error, most often a reversal of cause and effect ("the giraffe's neck grew longer because it kept stretching"), insulin and glucagon swapped, or mass "lost" in burning; 33 of the 41 had been copied without a word changed. The learners had not made these errors; they had failed to catch them.

That failure has a name. Automation bias is the use of an automated aid's output "as a heuristic replacement for vigilant information seeking and processing" (Mosier & Skitka, 1996, as cited in Parasuraman & Manzey, 2010, p. 391), and a chatbot's fluent, confident answer invites it. In a study of nearly 1,000 high school learners in Turkey, those given an unguarded GPT-4 chat interface for mathematics practice scored 48% higher on the practice problems but 17% lower than learners who never had it on an exam taken afterward without the tool, and the authors found that "students often use GPT Base as a 'crutch' by asking for and copying solutions" (Bastani et al., 2025, pp. 1–2). The same authors note that "generative AI is fallible and users must check its outputs" (p. 1). The question for a Science teacher is how to teach the checking.

Three lines of research shaped our routine. First, checking is a skill that can be taught: in a district-wide field study, high school teachers taught six 50-minute lessons in "lateral reading", "leaving an unfamiliar website to search the open Web before investing attention in the site at hand", and their students "grew significantly in their ability to judge the credibility of digital content" compared with students in matched schools (Wineburg et al., 2022).

DepEd's pedagogy cards for Key Stage 3 include a strategy of the same kind, Triangulation, in which learners "assess three short sources" and fill a grid with what the sources agree on, where they conflict, and what they leave out (DepEd Bureau of Learning Delivery, 2026). Second, finding and fixing errors in an example can help learners learn the content itself: sixth graders who were prompted "to find, explain, and fix errors in decimal problems" did better a week later than those who solved and explained the same problems (McLaren et al., 2015, p. 520). Third, and most important for how we measured: a lesson in doubt can make learners doubt everything. A media literacy intervention with adults in the United States and India "reduced the perceived accuracy of both mainstream and false news headlines," though it improved discernment because its effect on false headlines was larger (Guess et al., 2020, p. 15536), and a reanalysis of five studies of two online games found that, when comparable true and false items were used, the games "did not improve discrimination, but rather elicited more 'false' responses to all news items" (Modirrousta-Galian & Higham, 2023). Signal detection theory separates the two: the ability to tell true from false (discrimination) and the general tendency to call things false (response bias) are "conceptually independent," so "any given factor can influence either d′ or c or both" (Batailler et al., 2022, p. 83). A routine that only taught suspicion would raise the number of errors learners find and the number of correct explanations they reject; our measure had to tell the two apart.

The routine also needed a way to show learners where to look. The WikiProject AI Cleanup's advice page "Signs of AI writing", a field guide used

by Wikipedia's editors, lists patterns "typical of AI chatbots" with real examples. Some are about content and fit any explanation: "LLM writing often puffs up the importance of the subject matter," AI chatbots "tend to insert superficial analysis of information," "tend to attribute opinions or claims to some vague authority," and "omit specific, unusual, nuanced facts (which are statistically rare) and replace them with more generic, positive descriptions" ("Wikipedia:Signs of AI Writing," 2026). The page itself warns that "Not all text featuring these indicators is AI-generated" and that "The patterns here are also only potential signs of a problem, not the problem itself" ("Wikipedia:Signs of AI Writing," 2026). We took four of these content signs as places to check, never as a verdict, and left out the signs of style, such as punctuation and favorite words, which say nothing about whether a claim is true. Our learners were never asked whether a text was written by AI: every explanation they judged was labeled as AI-written.

The studies above come from the United States, India, and Turkey, with adults or with other subjects. Our own term's competencies call for a routine like this: three of the fourteen Term 1 competencies of Grade 10 Science ask learners to "use information from secondary sources" (DepEd, 2026b), and a chatbot's answer is now the secondary source most of our learners reach first. This study asked whether six short sessions of a checking routine would make Grade 10 learners better at telling correct AI-written explanations from wrong ones, without making them reject correct ones.

Innovation, Intervention, and Strategy

The innovation was "Check Before You Trust", a 20-minute routine in which learners judged a short explanation that an AI tool had written for the week's competency, checked it against sources they could trust, and fixed what was wrong.

The Explanations

The first researcher wrote the explanations with an AI tool, named in the declaration in Appendix F, before the term began. For each competency, the prompt asked for an explanation for a Grade 10 learner in 80 to 100 words. Each explanation was checked line by line against the reference notes and the curriculum, as DepEd requires of AI outputs teachers use: "strict verification of content accuracy to preclude sole reliance on AI-generated outputs" (DepEd, 2026a, Sec. VI.1.1). Half were then given one planted error, by editing one sentence; where the tool's text already held an error of its own, the error was kept as that explanation's one error or the sentence was corrected. Each explanation was also written in two wordings: plain, and flagged, with two or three phrases of the kinds the routine teaches learners to flag added and the science left the same. Every sheet said "Written by an AI tool, then checked and edited by your teacher." No learner's name, work, or other personal information was entered into the tool (Sec. VI.1.3), and the prompts and the chat links were kept (Sec. VI.1.7).

The Four Flags

Learners flagged, with a red pen, four kinds of wording that claim something without showing it. The four were written in our own words from

the content signs on the WikiProject AI Cleanup page as it stood on February 20, 2026 ("Wikipedia:Signs of AI Writing," 2026):

1. **A claim of importance with no fact that shows it:** "plays a crucial role", "is a testament to", "underscores the significance of".
2. **A tail that sounds like analysis but adds no step of cause and effect:** "…, highlighting the balance of nature", "…, ensuring stability".
3. **An authority with no name:** "scientists agree", "studies show", "experts say".
4. **A general word where a specific fact belongs:** "a substance", "an organ", "certain conditions", "many factors".

The rule printed on every sheet was "A flag tells you where to check. It does not tell you the answer."

The Routine

Each session took 20 minutes of the 50-minute Science period, on the Wednesday and Thursday of the cycle's three weeks, and used the week's competency:

1. **Read (2 minutes):** learners read the explanation once, without marking it.
2. **Flag (3 minutes):** learners underlined any of the four flags in red.
3. **Triangulate (8 minutes):** learners compared each claim of the explanation with two other sources, the reference notes and the lesson's own activity or diagram, and filled a grid with three columns, as in DepEd's Triangulation card: where the sources agree, where they

conflict, and what the explanation leaves out (DepEd Bureau of Learning Delivery, 2026).

4. **Decide and fix (4 minutes):** learners marked the explanation "correct" or "has an error"; for an error, they underlined the sentence and wrote the corrected sentence with the page of the notes that supports it.
5. **Debrief (3 minutes):** the teacher showed the key, the class named which flags had led to an error and which had led nowhere, and the teacher modeled checking one claim aloud.

Learners worked alone in steps 1 to 4. Two explanations were judged in each session, twelve in a cycle.

The Reference Notes

The two researchers wrote 12 pages of reference notes, one page for each of the twelve competencies the checks covered, from the curriculum's competencies and our lessons. Each learner kept a printed set in a plastic envelope. Every planted error in the practice explanations and in the tests could be corrected from the notes, so the routine and the tests asked learners to check, not to remember. The notes were written by people, as DepEd's guidelines ask of learning resources, which are to remain "primarily human-created" (DepEd, 2026a, Sec. V).

The Comparison

While one wave had a cycle, the other wave's sections spent the same 20 minutes of the same two periods on the practice exercises of the same competency from our lesson plans: problems, a diagram to complete, or

questions on the lesson's activity. Both waves had the same lessons, the same notes, and the same time on each competency.

Strategy and Evaluation Plan

Four criteria for success were set in the proposal: (a) after Cycle 1, learners who had the routine discriminate better than learners who had not yet had it, by at least 0.40 in d' and significant at the .05 level; (b) the routine is run as planned, with at least 90% of the steps done in the sessions the Head Teacher observed; (c) the routine does not make learners reject correct explanations because of their wording: after Cycle 1, the style gap (the false-alarm rate on flagged correct explanations less that on plain ones) of learners who had the routine is no more than 5 percentage points larger than that of learners who had not; and (d) learners who had the routine do not score more than 5 percentage points lower on the second summative test. The proposal also fixed what would happen if (c) failed in Cycle 1: the practice set would be revised before Cycle 2, and the revision decided only from the T2 results and recorded before Cycle 2 began.

Action Research Questions

Main question: Does the "Check Before You Trust" routine help Grade 10 learners tell correct AI-written science explanations from those with an error, without making them reject correct ones?

1. How well did the learners tell correct explanations from explanations with an error before the routine, in hits, false alarms, sensitivity (d'), and criterion (c)?

2. After Cycle 1, did the learners who had the routine discriminate better than the learners who had not yet had it?
3. Did Cycle 1 make the learners more likely to reject correct explanations because of their flagged wording?
4. Did Cycle 2 replicate the gain: did the second wave's sensitivity rise from T2 to T3 more than the first wave's?
5. After the revised Cycle 2, was the learners' tendency to reject correct explanations for their wording smaller than after Cycle 1?
6. Did learners who had the routine score differently on the second summative test from learners who had not yet had it?
7. How well did the learners locate and fix the errors they found, and did the first wave keep its gain 34 days after its last session?

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

$H_0$1: After Cycle 1, the sensitivity (d') of learners who had the routine does not differ from that of learners who had not yet had it.

$H_0$2: After Cycle 1, the style gap (the false-alarm rate on flagged correct explanations less that on plain ones) does not differ between learners who had the routine and learners who had not yet had it.

$H_0$3: From T2 to T3, the change in sensitivity of the second wave, which had Cycle 2, does not differ from that of the first wave, which had no new sessions.

$H_0$4: The style gap right after the revised Cycle 2 does not differ from the style gap right after Cycle 1.

$H_0$5: The second summative test scores of learners who had the routine do not differ from those of learners who had not yet had it.

Action Research Methods

Research Design

The study used a switching-replications design across the two researchers' eight Grade 10 sections. In this design, "the researcher administers treatment at a later date to the group that initially served as a no-treatment control" (Shadish et al., 2002, p. 146), so that the second group's treatment replicates the first group's. For each researcher, the school head drew on June 25, 2026, after the first check, one of six slips from a box, each slip naming two of the researcher's four sections: the two named began the routine in Cycle 1 (Wave 1), and the other two began it in Cycle 2 (Wave 2). The slips drawn were BD for the first researcher, giving the routine first to Darwin and Mendel, and AC for the second researcher, giving it first to Galileo and Newton. Drawing within each researcher kept each teacher's sections in both waves. There were 36 possible lots, six for each researcher.

Learners were tested three times, each time with one of three parallel forms: T1 on June 24, 2026, before either cycle; T2 on July 22, 2026, after Cycle 1 (July 1 to 16, 2026); and T3 on August 19, 2026, after Cycle 2 (July 29 to August 13, 2026). Each check came 6 days after the last session of the cycle before it. The study ran in Term 1 of School Year 2026–2027, whose Instructional Block ran from June 15 to September 1, 2026, with the second summative test on July 28, 2026 and the Term 1 Examination on August 28 and September 1, 2026 (DepEd, 2026c, Annex B).

Participants and Other Sources of Data and Information

The participants were the learners of the two researchers' eight Grade 10 sections in a public secondary school of 2,148 learners, School Year 2026–2027. All 356 enrolled learners were invited; the parents of 337 (94.7%) consented, and these learners gave their assent (Table 1). Every learner took part in every session of the cycle the lot gave the section; consent decided only whose papers and scores were recorded. Two learners with consent transferred to other schools during the term; their papers up to then were analyzed.

Table 1

Learners Enrolled and Consenting, by Section and Wave

Section	Teacher	Wave	Enrolled	Parents declined	Form not returned	With consent
Curie	First researcher	Wave 2	45	2	1	42
Darwin	First researcher	Wave 1	43	1	1	41
Faraday	First researcher	Wave 2	46	2	0	44
Mendel	First researcher	Wave 1	44	1	1	42
Galileo	Second researcher	Wave 1	44	1	1	42
Kepler	Second researcher	Wave 2	47	3	0	44
Newton	Second researcher	Wave 1	42	1	1	40
Pasteur	Second researcher	Wave 2	45	2	1	42

Note. Wave 1 had the routine in Cycle 1 (July 1 to 16, 2026); Wave 2 in Cycle 2 (July 29 to August 13, 2026). Learners present at the checks: 322 at T1, 321 at T2, and 322 at T3; 310 took both T1 and T2, and 306 both T2 and T3.

Sample Size

With eight whole sections, the number of learners was fixed. Before the study, we simulated 500 studies of this design for each of three differences,

with a standard deviation of about 0.8 in *d′*, as in the pilot (0.81), and an assumed correlation of .45 between two forms (the pilot gave one form, whose split-half reliability was .52). With the analysis planned below, the power to detect a difference of 0.40 in *d′* was .88, and .67 for 0.30. A difference of 0.40 is about what separates a learner who finds 40% of the errors from one who finds 56%, with the same false alarms.

Data Gathering Methods

The AI Explanation Check

The main instrument was a test of our own, the AI Explanation Check, in three parallel forms (A, B, and C) of 16 items each. Each item was one explanation of 80 to 100 words on one of twelve of the fourteen Term 1 competencies: we left out the competency on discussing the theories of evolution in small groups, a classroom activity, and joined the description of homeostasis to the competency on its feedback mechanisms. On each form, eight explanations were correct and eight had one planted error, and in each half, four were in flagged wording and four in plain wording. The eight errors on each form were of four kinds, two of each: a wrong term or value in place of the right one, a cause and effect reversed, mass or atoms created or destroyed, and a rule stretched past where it holds. For each item, the learner marked "correct" or "has an error"; for an error, underlined the wrong sentence and wrote the corrected sentence; and ticked one reason: "I found it in the notes (page __)", "The wording made me doubt it", "It sounded right (or wrong) to me", or "I guessed". Learners had their reference notes open and no device; each check took 30 minutes.

Every learner took each form once. The three forms were given in one of six orders (ABC, ACB, and so on), dealt down a shuffled class list in each section, so that each form was used about equally at each check and in each wave.

The first researcher drafted 56 items in March 2026, and a panel of five validators judged each one: whether the explanation was correct or wrong as keyed, whether the error could be found and corrected from the notes, and whether its two wordings said the same science. Lawshe's content validity ratio, CVR = $(n_e - N/2)/(N/2)$, was computed for each item (Lawshe, 1975, p. 567); with a panel of five, the critical value at the .05 level is 1.00, so an item was kept only when all five agreed (Ayre & Scally, 2014, p. 82). Seven items were revised and judged again, and eight were dropped, leaving 48. Form A was tried on March 17, 2026 with one Grade 10 section of the previous school year (46 learners, no names recorded): the median learner finished in 24 minutes and the slowest in 29, so 30 minutes were allowed; three items that learners found unclear were reworded; and the split-half reliability of d' was .52.

Scoring

A verdict was scored against the key. For each learner and check, a hit was an explanation with an error marked "has an error" (out of 8) and a false alarm was a correct explanation marked "has an error" (out of 8). For each error found, the scorer recorded whether the learner had underlined the right sentence (located) and whether the corrected sentence was right (fixed), against a key the validators had approved. A second Science teacher scored

20% of the papers again, drawn by lot by code, without knowing the learner's section, wave, or check.

Other Sources

The second summative test of Term 1 (25 items, department-made, on July 28, 2026) and each learner's final grade in Grade 9 Science came from the class records, with consent. On that date Wave 1 had finished Cycle 1 and Wave 2 had not begun Cycle 2. Each teacher logged the minutes and attendance of every session. The Head Teacher observed two sessions in each section, drawn by lot, with a checklist of the five steps.

Storage

Papers were stored by code in a locked cabinet in the Science faculty room. The key that links codes to names was kept by the school's research coordinator, apart from the papers, and will be destroyed when the report is accepted.

Ethical Considerations

Parents received a letter in Filipino and English at the parents' orientation on June 10, 2026, in the Opening Block, with the school's notice of the AI tool the teachers would use to prepare materials, its risks, and the safeguards, as DepEd asks teachers to give parents "At the commencement of each school year" (DepEd, 2026a, Sec. VI.2.5). Learners whose parents consented signed an assent form on June 18, 2026, before the first check, as DepEd's quality tools for completed research ask (DepEd, 2022, Annex 2, Item 13); the class adviser, who was not on the research team, witnessed each one.

Learners without consent had the same sessions and took the checks as class work; their papers were not recorded.

The learners never used an AI tool in the study: not in the sessions, and not in the checks, which were supervised assessments of the kind DepEd keeps "AI-prohibited" (DepEd, 2026d, para. 35). Only the first researcher used one, to draft the explanations the learners judged, in the practice sets and in the checks. DepEd lets teachers use AI "in the creation and enhancement of instructional materials, subject to existing quality assurance and approval processes" (DepEd, 2026a, Sec. VI.1.1(ii)) and to develop assessment tools, provided the process is not "fully automated or solely AI-generated" (Sec. VI.1.2(i); DepEd, 2026d, para. 35). DO 003 also limits AI in research work to "ancillary support functions, such as the analysis of processed data, grammatical refinement, and citation verification" (DepEd, 2026a, Sec. VI.1.3), and it does not say whether material drafted for a research instrument counts as such work. Every key, planted error, validation, and analysis here was ours, and the declaration in Appendix F names the tool and its uses. The checks were formative and were not graded: "Formative assessment shall be used to monitor learning progress but shall not be used for grade computation" (DepEd, 2026d, para. 15). A learner's answers and grades are personal information, and information about a learner's education is sensitive personal information (Republic Act No. 10173, 2012, Secs. 3(g) and 3(l)), so the data set holds codes, and no personal information entered an AI system (DepEd, 2026a, Sec. VI.1.3).

Data Analysis

For each learner and check, the hit rate H and the false-alarm rate F gave sensitivity, $d' = \Phi^{-1}(H) - \Phi^{-1}(F)$, and criterion, $c = -[\Phi^{-1}(H) + \Phi^{-1}(F)]/2$, where Φ^{-1} turns a proportion into a z score (Stanislaw & Todorov, 1999, p. 142). A d' of 0 means a learner could not tell correct from wrong explanations at all; a c above 0 means a lean toward calling explanations correct. Because a rate of 0 or 1 has no finite z score, every count was corrected by the loglinear rule, "adding 0.5 to both the number of hits and the number of false alarms and adding 1 to both the number of signal trials and the number of noise trials" (Stanislaw & Todorov, 1999, p. 144; Hautus, 1995). The style gap was the false-alarm rate on the four flagged correct explanations less that on the four plain ones, in percentage points.

$H_0$1, $H_0$2, $H_0$4, and $H_0$5 were tested with linear mixed models with a random intercept for section, the outcome at the later check as the dependent variable, the same outcome at T1 (for $H_0$5, the Grade 9 Science grade) as a covariate, and the teacher and the form as fixed effects. Because the lot assigned whole sections, the test of the wave used t with the between-within degrees of freedom: eight sections less three section-level terms (intercept, wave, teacher), $df = 5$. In simulations of cluster-randomized trials with binary outcomes, this between-within method "maintains the nominal type I error rates even when the total number of clusters is as low as 10" (Li & Redden, 2015, p. 1); our outcomes are continuous, and our eight sections are fewer than any number they tested. For $H_0$3, the waves were not equal at T2 by design, so the outcome was each learner's change in d' from T2 to T3, with

the teacher and both forms as fixed effects. $H_0 1$ was the primary hypothesis; $H_0 2$ to $H_0 5$ were adjusted for four tests by Holm's (1979) procedure. A randomization test asked how the drawn lot ranked among all 36 lots the boxes could have given, using each section's mean gain in d' from T1 to T2.

Three further analyses checked the main result. A model of all three checks together, with a random intercept for each learner within section and fixed effects for the check, the teacher, and the form, estimated the routine's effect across both cycles, as in the analysis of a stepped-wedge trial (Hussey & Hughes, 2007, p. 184). A probit model of each verdict, fitted by generalized estimating equations with learners as clusters, estimated d' directly as the coefficient of "the explanation has an error" (DeCarlo, 1998, p. 190). The split-half reliability of d' (odd and even items, each half with two items of every kind, Spearman-Brown corrected) was computed at each check, and the agreement of the two scorers on located and fixed errors by Cohen's kappa (Cohen, 1960). The analyses were run in Python with statsmodels (Seabold & Perktold, 2010); the Excel workbook in Appendix A recomputes every score and every statistic that a spreadsheet formula can.

Discussion of Results and Reflection

Implementation

The 48 sessions were held as scheduled, 21.0 minutes on average (from 18 to 24). Learners attended 94.4% of their sessions, and 71.5% attended all six. In the 16 sessions the Head Teacher observed, 78 of the 80 steps were done; the debrief was cut short twice for lack of time. Criterion (b), at least 90% of the steps, was met.

Question 1: Before the Routine

At T1, the learners accepted 56.0% of the explanations with an error as correct (1,443 of 2,576 verdicts): the mean hit rate was 44.0%, and the mean false-alarm rate 23.1% (Table 2). Mean sensitivity was low, d' = 0.59 (*SD* = 0.89), and the criterion leaned toward trust, c = 0.46. The most common reason ticked was "It sounded right (or wrong) to me" (58.9% of verdicts); learners wrote a page of the notes for 6.0% of their verdicts, though the notes were on their desks. Of the errors they did find, they located 67.2% and fixed 39.2%. The two waves did not differ at T1: d' = 0.56 in Wave 1 and 0.61 in Wave 2.

Table 2

Hits, False Alarms, Sensitivity, Criterion, and Style Gap by Wave and Check

Wave and check	*n*	Hit rate (%)	False-alarm rate (%)	d' *M* (*SD*)	*c M* (*SD*)	Style gap (points)
Wave 1, T1	155	43.1	23.3	0.56 (0.88)	0.47 (0.49)	7.3
Wave 2, T1	167	44.8	22.8	0.61 (0.90)	0.44 (0.43)	4.6
Wave 1, T2 (after Cycle 1)	161	72.4	25.0	1.33 (0.96)	0.03 (0.46)	18.9
Wave 2, T2 (waiting)	160	45.2	20.9	0.67 (0.94)	0.47 (0.40)	6.1
Wave 1, T3 (34 days after Cycle 1)	155	72.3	25.4	1.28 (0.82)	0.04 (0.44)	15.3
Wave 2, T3 (after Cycle 2)	167	73.1	21.0	1.46 (0.90)	0.09 (0.42)	6.3

Note. Hit rate: explanations with an error marked "has an error", of 8. False-alarm rate: correct explanations marked "has an error", of 8. d' and c with the loglinear correction. Style gap: the false-alarm rate on flagged correct explanations less that on plain ones, of 4 each, in percentage points, computed before rounding. Learners present at each check.

Question 2: After Cycle 1

At T2, learners who had had Cycle 1 discriminated better than learners still waiting: adjusted for T1, the difference in d' was 0.69, 95% CI [0.45, 0.93],

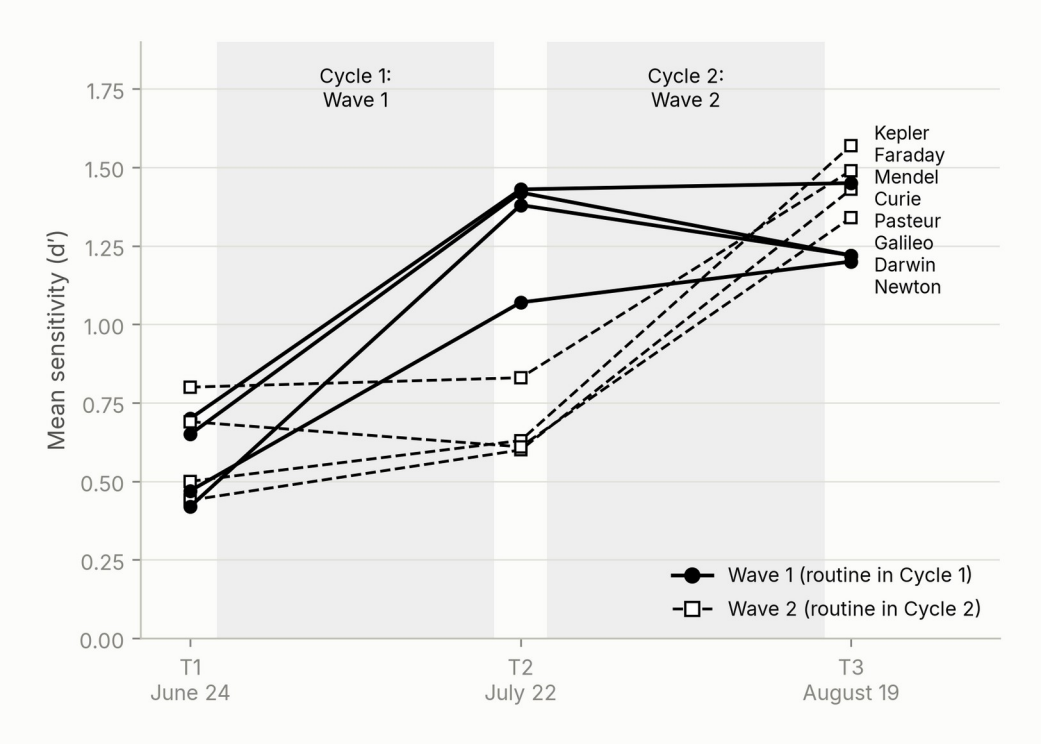
$t(5) = 7.45$, $p < .001$, Hedges' $g = 0.71$ (Table 3). $H_0 1$ is rejected, and the difference met criterion (a), 0.40. The gain came from hits: the hit rate rose by 28.0 percentage points more in Wave 1, 95% CI [22.0, 33.9], while the false-alarm rate rose by 3.8 points more in Wave 1, 95% CI [−2.9, 10.5], $p = .206$. The criterion moved from trust toward doubt, by −0.45, 95% CI [−0.58, −0.31].

Each section's mean gain from T1 to T2 was, in order: Darwin 0.97, Mendel 0.74, Galileo 0.73, Newton 0.59, Curie 0.16, Kepler 0.13, Faraday 0.02, Pasteur −0.12 (Figure 1). Every Wave 1 section gained more than any Wave 2 section, so the lot the school head drew gave the largest difference between the waves of all 36 lots the two boxes could have given: a randomization test gives $p = .028$ (1/36) for the direction the proposal named and $p = .056$ (2/36) in either direction. With 36 lots, no result could have given a two-sided p below 2/36.

The gain was not limited to the competencies Cycle 1 had practiced (chemical equations, conservation of mass, and balancing equations). At T2, Wave 1 judged 71.7% of the items on those competencies correctly against 65.3% in Wave 2, and 74.4% of the items on the other nine competencies against 61.2%.

Figure 1

Each Section's Mean Sensitivity (d') at the Three Checks



Note. Solid lines: Wave 1 (Cycle 1, July 1 to 16, 2026). Dashed lines: Wave 2 (Cycle 2, July 29 to August 13, 2026). Learners present at each check.

Question 3: The Style Gap After Cycle 1

Cycle 1 made learners more likely to reject a correct explanation because of its wording. Before the routine, the style gap was 7.3 points in Wave 1 and 4.6 in Wave 2; at T2 it was 18.9 points in Wave 1 and 6.1 in Wave 2, an adjusted difference of 12.9 points, 95% CI [4.5, 21.3], Holm-adjusted p = .029. $H_0$2 is rejected, and criterion (c) was not met. Of the 222 false alarms on flagged correct explanations in Wave 1 at T2, 47.3% were ticked "The wording made me doubt it", against 26.4% before the routine (both waves at T1 and Wave 2 at T2).

The cause was in our practice set. In Cycle 1's twelve practice explanations, five of the six in flagged wording had an error and one of the six

in plain wording had one: without meaning to, we had taught that flags meant errors. As the proposal planned, we revised the set on July 24, 2026, before Cycle 2: in the new set, three of the six flagged explanations had an error and three did not, and the same held for the plain ones. The debrief gained a question, "Which flags led nowhere today?", and the rule on the sheet was read aloud at every session.

Question 4: The Replication in Cycle 2

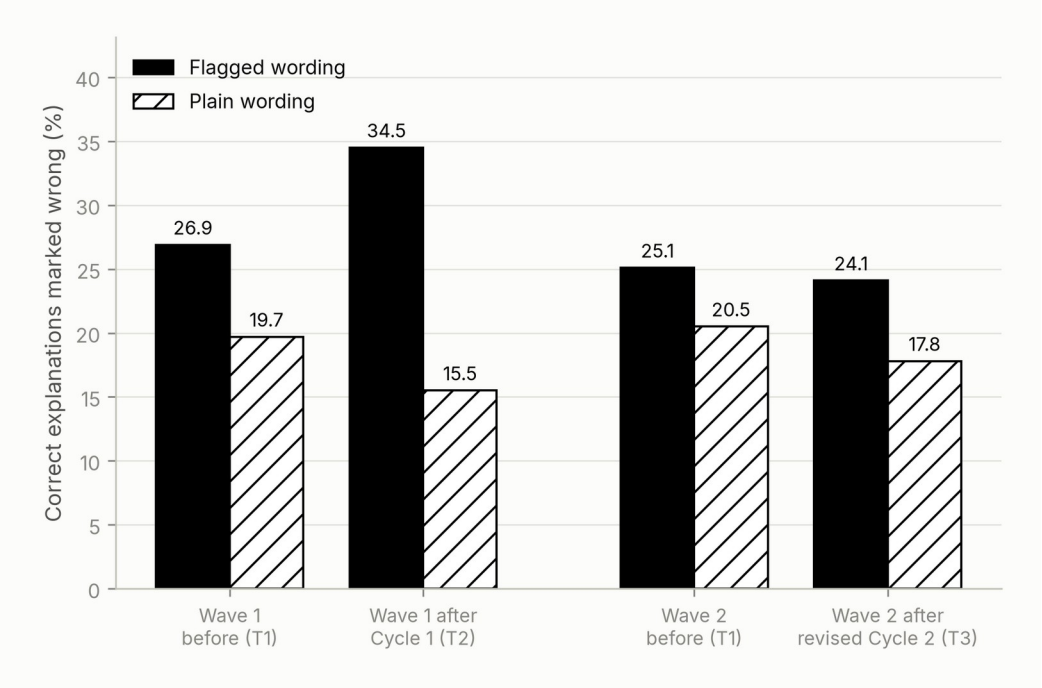
From T2 to T3, among the 306 learners present at both checks, Wave 2's mean d' changed by +0.79 with Cycle 2, while Wave 1's changed by −0.12 with no new sessions: a difference of 0.91, 95% CI [0.61, 1.21], Holm-adjusted $p = .002$. $H_0 3$ is rejected: the gain was replicated in the other four sections, with the revised practice set. From each wave's own T1, Wave 1's d' rose by 0.76 by T2 and Wave 2's by 0.86 by T3.

Question 5: The Style Gap After Cycle 2

Right after the revised Cycle 2, the style gap in Wave 2 was 6.3 points, against 18.9 points in Wave 1 right after Cycle 1: adjusted for T1, a difference of −12.7 points, 95% CI [−20.8, −4.7], Holm-adjusted $p = .029$ (Figure 2). $H_0 4$ is rejected. Of Wave 2's false alarms on flagged correct explanations at T3, 15.5% were ticked "The wording made me doubt it". Wave 1's gap, which no revision reached, was still 15.3 points at T3.

Figure 2

False-Alarm Rates on Correct Explanations, Flagged and Plain, Before and After Each Cycle



Note. The share of correct explanations marked "has an error", by wording. Before: each wave at T1. After: Wave 1 at T2, Wave 2 at T3.

Question 6: The Second Summative Test

On the second summative test of July 28, 2026, Wave 1, which had finished Cycle 1, scored 60.3% (*SD* = 14.7), and Wave 2, which had not begun Cycle 2, 63.2% (*SD* = 15.8). Adjusted for the Grade 9 Science grade, the difference was −1.9 points, 95% CI [−7.6, 3.8], Holm-adjusted $p = .438$. $H_0 5$ is not rejected. The interval, though, reaches a drop of more than 5 points, so the study cannot confirm criterion (d): a small cost to the summative test is not ruled out.

Question 7: Locating, Fixing, and Keeping

Of the errors learners found, they located 67.9% and fixed 39.0% before the routine (each wave's checks before its cycle: both waves at T1 and Wave

2 at T2), 88.7% and 74.4% right after Cycle 1, and 88.4% and 71.2% right after Cycle 2. The share of verdicts with a page of the notes written rose from 6.1% before the routine to 74.5% after Cycle 1 and 81.3% after Cycle 2. The two scorers agreed on 93.9% of the located judgments, $\kappa = .82$, and on 91.7% of the fixed judgments, $\kappa = .83$ (964 errors found on 188 papers).

Wave 1 kept most of its gain. From T2 to T3, 34 days after its last session, the 151 Wave 1 learners present at both checks moved from a mean d' of 1.385 to 1.266, a change of −0.12, 95% CI [−0.26, 0.02], $t(150) = -1.67$, $p = .096$: 86% of the gain over T1 remained.

Other Checks

Table 3

Tests of the Five Null Hypotheses

Hypothesis	Outcome	Estimate	95% CI	$t(5)$	p	Holm p	Decision
$H_0$1	d' at T2, Wave 1 less Wave 2	0.69	[0.45, 0.93]	7.45	< .001	(primary)	Rejected
$H_0$2	Style gap at T2 (points), Wave 1 less Wave 2	12.9	[4.5, 21.3]	3.95	.011	.029	Rejected
$H_0$3	Change in d' from T2 to T3, Wave 2 less Wave 1	0.91	[0.61, 1.21]	7.77	< .001	.002	Rejected
$H_0$4	Style gap after the cycle (points), Cycle 2 less Cycle 1	−12.7	[−20.8, −4.7]	−4.07	.010	.029	Rejected
$H_0$5	Second summative test (%), Wave 1 less Wave 2	−1.9	[−7.6, 3.8]	−0.84	.438	.438	Not rejected

Note. Linear mixed models with a random intercept for section; the outcome at T1 (for $H_0$5, the Grade 9 Science grade), the teacher, and the form as fixed effects; for $H_0$3, the change score with the teacher and both forms. t with the between-within degrees of freedom (eight sections less three section-level terms). Holm's adjustment over $H_0$2 to $H_0$5.

In the models of T2, the section variance was estimated at zero: the eight sections differed no more than their learners did, and the t test on five

degrees of freedom is what allows for the sections. A *t* test on the eight sections' mean gains from T1 to T2 gave the same answer, $t(6) = 7.10$, $p < .001$. The model of all three checks together put the routine's effect at 0.75 in *d′*, 95% CI [0.59, 0.91] (965 papers of 337 learners). The probit model of each of 15,440 verdicts estimated *d′* before the routine at 0.62, 95% CI [0.52, 0.72], and its rise with the routine at 0.72, 95% CI [0.62, 0.82], the same answer by a different route. The three forms gave similar *d′* at T1 (A: 0.55, B: 0.52, C: 0.69; $F(2, 319) = 1.07$, $p = .345$). The split-half reliability of *d′* was .61 at T1, .67 at T2, and .55 at T3: enough to compare groups of learners, not to judge one learner by one form.

Reflection

What We Found

Six sessions of twenty minutes made our learners better at telling correct AI-written explanations from wrong ones, by 0.69 in *d′*, and the gain was replicated in the other four sections. Before the routine, a learner given an explanation with an error accepted it more often than not, and most said it "sounded right". After it, learners found most errors, wrote where in the notes they had checked, and fixed most of what they found. The gain was at least as large on competencies the cycle had not practiced as on those it had, which suggests the learners learned to check, not only the content of the practice.

The routine also taught something we did not intend. After Cycle 1, learners rejected correct explanations more often when their wording was flagged, and nearly half of those rejections were ticked "The wording made me

doubt it". The cause was our practice set, in which five of six flagged explanations had an error. When the set was balanced for Cycle 2, the gain in d' came again and the style gap mostly did not. This is the warning of the media literacy research (Guess et al., 2020; Modirrousta-Galian & Higham, 2023) in our own classrooms: what a lesson in checking teaches depends on the examples it practices on. It is also what the WikiProject AI Cleanup page says of its signs: they are "only potential signs of a problem, not the problem itself" ("Wikipedia:Signs of AI Writing," 2026).

How It Compares With the Research

Our finding agrees with the high school lateral reading study (Wineburg et al., 2022) that checking against other sources can be taught in a few short lessons, and with Batailler et al. (2022) that an intervention's effect on discrimination and on bias must be measured apart: our learners' criterion moved toward doubt by −0.45 while their sensitivity rose, and a count of errors found alone would have shown only the gain. Unlike the learners who used an unguarded chatbot for practice and did worse later without it (Bastani et al., 2025), our learners never used the tool; they judged its output, and the skill they practiced was the one the tool cannot supply.

What Else Could Explain the Difference

Taking the check a second time cannot explain the gain: both waves took the same checks on the same days, and Wave 2's d' rose only slightly from T1 to T2 (0.61 to 0.67). The forms cannot explain it: each form was used at each check in each wave, and the forms were alike at T1. Knowledge of the content cannot fully explain it, since the notes were open at every check and

the gain was at least as large on competencies Cycle 1 had not practiced. The teachers knew which sections had the routine, and they taught both; but the verdicts were scored against a key, and the located and fixed errors were scored again by a teacher who did not know the wave. Eight sections are few: the randomization test could not reach a two-sided *p* below .056, and the models' tests rest on five degrees of freedom. The replication in the other four sections, with the same result, is the stronger part of the evidence.

Limits

The study ran in eight sections of one school, in one term. The explanations were written by one AI tool and checked and edited by a teacher; the flagged wording was added by us, so the test measured whether learners could see past wording of the kinds we chose, not every kind an AI tool produces. Learners checked against notes we wrote, which held every correction; outside school, a learner must find a source, judge it, and may find none. The checks were short, and their reliability (.55 to .67) is too low to judge one learner. The revision of the practice set came with Cycle 2, so Cycle 2's smaller style gap may owe something to its later dates and to the debrief question as well as to the balanced set; as Shadish et al. (2002) note of this design, "the second phase is not an exact replication" (p. 146). The second summative test could not rule out a small cost. And the routine did not make learners immune to a confident answer: automation bias, a review of the research concluded, "cannot be prevented by training or instructions" (Parasuraman & Manzey, 2010, p. 381); what the routine changed was what learners did before they trusted.

What We Would Change

We would balance the practice set from the first session, with as many flagged correct explanations as flagged wrong ones; lengthen each form to 24 items to raise its reliability; and add, in a later cycle, explanations whose errors the notes do not settle, so that learners practice saying "I cannot tell yet" and finding a source.

Conclusion

The "Check Before You Trust" routine helped Grade 10 learners tell correct AI-written science explanations from wrong ones, in two cycles in eight sections. Its flags help learners find where to check only when the practice teaches that a flag is not a verdict. We recommend it for Grade 10 Science, with a practice set in which flags and errors are not paired.

Action Plan

Table 4

Action Plan

Activity	Objective	Time frame	Persons involved	Resources	Expected output
Share the routine and the revised practice set at the Science department's learning action cell	Let the department's Grade 10 teachers use it in Term 2	October 2026	Researchers, Head Teacher	Practice sheets, reference notes	The routine used in every Grade 10 section
Write practice sets and reference notes for the Term 2 and Term 3 competencies, flags balanced	Keep the routine running for the school year	October to December 2026	Researchers, Master Teachers	Curriculum and budget of work	Two balanced sets, each validated by the panel
Lengthen the AI Explanation Check to 24 items a form	Raise reliability toward .75	November 2026 to January 2027	Researchers, validators	Item bank	Three validated 24-item forms
Try a "cannot tell yet" step with sources from the school library	Practice judging when the notes do not settle a claim	Term 3, January to March 2027	Researchers, school librarian	Library books on the term's topics	A short trial with one section, reported to the department
Present the results at the division research forum and submit the report for the division's repository	Share the findings	As the division schedules	Researchers	Report and slides	Presentation and archived report
Give the parents' notice on AI in instruction at the start of School Year 2027–2028	Meet DepEd's yearly notice	June 2027	Researchers, advisers	Notice in Filipino and English	Notice given and filed

References

Ayre, C., & Scally, A. J. (2014). Critical values for Lawshe's content validity ratio: Revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development, 47*(1), 79–86. https://doi.org/10.1177/0748175613513808

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences, 122*(26), Article e2422633122. https://doi.org/10.1073/pnas.2422633122

Batailler, C., Brannon, S. M., Teas, P. E., & Gawronski, B. (2022). A signal detection approach to understanding the identification of fake news. *Perspectives on Psychological Science, 17*(1), 78–98. https://doi.org/10.1177/1745691620986135

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods, 3*(2), 186–205. https://doi.org/10.1037/1082-989X.3.2.186

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. (2022). *Adopting tools to improve quality management of completed research at the Department of Education* (DepEd Memorandum No. 028, s. 2022). https://www.deped.gov.ph/2022/04/05/april-5-2022-dm-028-s-2022-adopting-tools-to-improve-quality-management-of-completed-research-at-the-department-of-education/

Department of Education. (2026a). *Foundational guidelines on artificial intelligence (AI) 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/

Department of Education. (2026b). *Budget of work: Science, Grade 10, School Year 2026–2027*. Learning Systems Strand. https://sites.google.com/deped.gov.ph/lsguide/budgets-of-work

Department of Education. (2026c). *Guidelines on the implementation of the three-term school calendar in basic education* (DepEd Order No. 009, s. 2026). https://www.deped.gov.ph/2026/04/16/april-16-2026-do-009-s-2026-guidelines-on-the-implementation-of-the-three-term-school-calendar-in-basic-education/

Department of Education. (2026d). *Revised guidelines on classroom assessment, grading system, and awards and recognition for the K to 12 Basic Education Program* (DepEd Order No. 015, s. 2026). https://www.deped.gov.ph/2026/06/04/june-4-2026-do-015-s-2026-revised-guidelines-on-classroom-assessment-grading-system-

and-awards-and-recognition-for-the-k-to-12-basic-education-program/

Department of Education Bureau of Learning Delivery. (2026). *Pedagogy cards* [Teaching strategies]. Learning Systems Strand. https://sites.google.com/deped.gov.ph/lsguide/lesson-planning

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences, 117*(27), 15536–15545. https://doi.org/10.1073/pnas.1920498117

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d′. *Behavior Research Methods, Instruments, & Computers, 27*(1), 46–51. https://doi.org/10.3758/BF03203619

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.

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

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Li, P., & Redden, D. T. (2015). Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analyzing

binary outcome in small sample cluster-randomized trials. *BMC Medical Research Methodology, 15*, Article 38. https://doi.org/10.1186/s12874-015-0026-x

McLaren, B. M., Adams, D. M., & Mayer, R. E. (2015). Delayed learning effects with erroneous examples: A study of learning decimals with a web-based tutor. *International Journal of Artificial Intelligence in Education, 25*(4), 520–542. https://doi.org/10.1007/s40593-015-0064-x

Modirrousta-Galian, A., & Higham, P. A. (2023). Gamified inoculation interventions do not improve discrimination between true and fake news: Reanalyzing existing research with receiver operating characteristic analysis. *Journal of Experimental Psychology: General, 152*(9), 2411–2437. https://doi.org/10.1037/xge0001395

Organisation for Economic Co-operation and Development. (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

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381–410. https://doi.org/10.1177/0018720810376055

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 *Proceedings of the 9th Python in Science Conference* (pp. 92–96).

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

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*(1), 137–149. https://doi.org/10.3758/BF03207704

Wikipedia:Signs of AI writing. (2026, February 19). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Wikipedia:Signs_of_AI_writing&oldid=1339272395

Wineburg, S., Breakstone, J., McGrew, S., Smith, M. D., & Ortega, T. (2022). Lateral reading on the open Internet: A district-wide field study in high school government classes. *Journal of Educational Psychology, 114*(5), 893–909. https://doi.org/10.1037/edu0000740

Financial Report

Table 5

Statement of Expenses

Date	Particulars	Receipt no.	Quantity	Amount (PhP)	Category
June 8, 2026	Parents' notice on AI in instruction, consent letters in Filipino and English, and assent forms	OR 3041	760 pages at 1.00	760.00	Reproduction
June 19, 2026	Reference notes, 12 pages, one set for each learner and spares	OR 3058	4,440 pages at 1.00	4,440.00	Reproduction
June 19, 2026	Clear plastic envelopes to keep each learner's notes	OR 3058	370 pieces at 8.00	2,960.00	Supplies and materials
June 22, 2026	Test forms A, B and C, 4 pages each, 370 copies of each	OR 3062	4,440 pages at 1.00	4,440.00	Reproduction
June 29, 2026	Practice sheets, Cycle 1: 6 sessions, 2 pages each, 190 copies	OR 3077	2,280 pages at 1.00	2,280.00	Reproduction
June 29, 2026	Red ballpoint pens for flagging, boxes of 50	OR 3077	8 boxes at 160.00	1,280.00	Supplies and materials
July 27, 2026	Practice sheets, Cycle 2 (revised set): 6 sessions, 2 pages each, 190 copies	OR 3104	2,280 pages at 1.00	2,280.00	Reproduction
October 16, 2026	Printing and binding of the final report	OR 3240	3 copies at 320.00	960.00	Reproduction
Total				**19,400.00**	

Note. Fund source: Basic Education Research Fund, school level, PhP 20,000.00, released PhP 16,000.00 with the memorandum of agreement and PhP 4,000.00 on the acceptance of the final report (DepEd, 2017, Sec. V.C.i.5). Unspent balance of PhP 600.00 returned. Every line is an expense the fund allows: supplies and materials, and reproduction, printing, and binding (Sec. V.C.i.4). Receipt numbers are invented for this sample.

Appendices

Appendix A. Data Set and Notebook

The Excel data set (the lot, the item map, every verdict with its located, fixed, page, and reason, the second scorer's judgments, the sessions, the fidelity checks, the second summative test, and every statistic Excel can compute as a formula, with Python's values for the rest, labeled) and the Jupyter notebook that reproduces every result from the data.

Appendix B. Parents' Notice, Consent Letter, and Assent Form

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

Appendix C. The AI Explanation Check

The three forms, the key, the validators' judgments, and the corrected sentences the scorers accepted. Kept with the school's copy so that the items can be used again.

Appendix D. The Routine

The four flags, the five steps with their minutes, the triangulation grid, the revised practice set's plan, and sample explanations with their keys (the kit published with this report).

Appendix E. Fidelity Checklist

The Head Teacher's checklist of the five steps, and the draw of the sessions observed.

Appendix F. Declaration of AI Use

We, the researchers, declare that the first researcher used [name and version of the AI tool] to write the first drafts of the explanations learners

judged in the practice sessions and the checks, for the purpose of preparing instructional and assessment materials. The prompt used for each competency was: "Explain [the competency] to a Grade 10 learner in 80 to 100 words." Every explanation was checked line by line against the reference notes and edited by the first researcher; the planted errors, the flagged wording, the keys, and every analysis are the researchers' own. We also used [name of the AI tool] to check the grammar of the English text of this report, with the prompt "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, codes, scores, or other personal information were entered into any AI tool (adapted from Annex A of DepEd Order No. 003, s. 2026).