Predict-Observe-Explain Lessons With Simulations in Grade 9 Science: A Crossover Study

An Action Research

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

Conducted June to July 2026

Abstract

This action research tested whether Predict-Observe-Explain (POE) lessons with computer simulations from the PhET Interactive Simulations project at the University of Colorado Boulder would raise Grade 9 learners' Science achievement and weaken their misconceptions about force and current. Two sections of a public secondary school took part in Term 1 of School Year 2026–2027 in a crossover design: Section Narra (n = 41) had POE lessons in Unit 1, Newton's laws of motion, and the regular lessons in Unit 2, electric current and circuits; Section Molave (n = 39) had the reverse, so every learner had both methods. Each unit had a 25-item pretest and posttest, validated by five experts and piloted with 40 Grade 10 learners, whose wrong options were written to carry known misconceptions. Posttest scores were analyzed with a linear mixed-effects model (method, unit, section, and pretest as fixed effects; learners as random effects), and misconception answers with a Poisson generalized estimating equation. Unit 2 was harder for both sections, by 10.0 percentage points. Adjusting for unit, order, and pretest, learners scored 5.7 points higher on the posttest of their POE unit, 95% CI [2.0, 9.3], p = .002 (d = 0.29 by the classical crossover test), and chose misconception answers 14% less often, rate ratio = 0.86, 95% CI [0.79, 0.92], p < .001. Seven of the eight misconceptions were chosen less after POE lessons. Learners said writing a prediction first made them watch closely. The school will use POE lessons where everyday misconceptions are common.

Keywords: predict-observe-explain, computer simulations, misconceptions, crossover design, linear mixed-effects model

Acknowledgment

The researchers thank the five experts who rated every draft item, two of whom also checked the misconception each wrong option was written to catch; the Grade 10 adviser and learners who answered the pilot tests during the Opening Block; the school's information and communications technology coordinator, who copied the offline simulations to the computer laboratory's laptops and kept them charged; the Schools Division Research Committee, whose reviewers asked for the pretest to be used in the analysis; the learners of Narra and Molave, who wrote their predictions before every simulation; and their parents, who gave their consent. The study was funded by the Basic Education Research Fund (BERF) for 2026.

Predict-Observe-Explain Lessons With Simulations in Grade 9 Science: A Crossover Study

Context and Rationale

Learners do not come to Science with empty heads. Before any lesson on force, most have learned from pushing carts and riding jeepneys that a moving thing stops when nothing pushes it, and from flashlights that a bulb "uses" electricity. Physics education research has documented these everyday ideas for decades. Halloun and Hestenes (1985) found that beginning physics students' common sense beliefs about motion, such as the belief that motion needs a force in its direction, were stable and changed little after conventional instruction; the Force Concept Inventory was built to detect them (Hestenes et al., 1992). For electric circuits, students commonly reason that current is used up by a bulb, that a battery gives the same current whatever is connected to it (McDermott & Shaffer, 1992), and that a change in a circuit affects only what comes after it (Shipstone, 1984).

Our school's own data show the same ideas. In March 2026, the science department gave a ten-item diagnostic of everyday ideas about force and electricity to the school's 83 Grade 8 learners, who would be the Grade 9 learners of this study. On the item about a cart rolling on a smooth floor after a push, 59% chose an answer saying it slows down because the push is used up; on the item about ammeters before and after a bulb, 66% chose an answer saying less current comes out of the bulb. The science teachers also noted why the regular hands-on activities sometimes leave these ideas in place: an improvised cart on the classroom floor does stop, which seems to confirm that

motion needs a force, and the laboratory's old dry cells and flashlight bulbs often give dim or uneven light, so learners see brightness that does not follow the rules they are taught.

Two lines of research suggest a remedy. In a Predict-Observe-Explain (POE) task, a learner first writes a prediction and the reason for it, then observes what happens, and then explains any difference between the two (White & Gunstone, 1992); committing to a prediction in writing brings the learner's own idea into the open, where the observation can test it. Computer simulations can show what an improvised setup hides: the PhET simulations of the University of Colorado Boulder let a learner turn friction off, show force and current that the eye cannot see, and give a battery whose voltage does not sag (Wieman et al., 2008). In a study of an introductory physics course, students who learned about circuits with the PhET Circuit Construction Kit instead of real equipment did better on conceptual questions about circuits and were faster at building a real circuit (Finkelstein et al., 2005), and a review of 51 studies found that simulations generally improved traditional science instruction, especially laboratory activities (Rutten et al., 2012).

Most of that evidence comes from universities and from countries where every learner has a computer. Whether POE lessons with simulations help Filipino Grade 9 learners in the first term of the new Science curriculum, with laptops shared by three learners, is a question for our own classroom. This action research compared POE lessons with simulations and the regular lessons in two units, in a design that gave every learner both methods. It

addresses the Teaching and Learning theme of the Basic Education Research Agenda (Department of Education [DepEd], 2016).

Innovation, Intervention, and Strategy

The innovation replaced the hands-on activity of a regular lesson with a Predict-Observe-Explain task done with a PhET simulation. Everything else in a lesson stayed the same: the competency, the 45 minutes, the teacher, the review, and the homework.

The Lesson Structure

Table 1

A Predict-Observe-Explain Lesson and a Regular Lesson

Part of the 45 minutes	POE lesson with a simulation	Regular lesson
Opening (5 minutes)	Review and the lesson's question	Review and the lesson's question
Activity (25 minutes)	Predict: each learner writes a prediction and a reason on the worksheet before touching the laptop (8 minutes). Observe: groups of two or three run the simulation and record what happens (10 minutes). Explain: each learner writes why the observation agrees or disagrees with the prediction (7 minutes)	The lesson exemplar's activity with real or improvised materials, such as carts, ramps, dry cells, bulbs, and wires, done in groups and recorded on the activity sheet
Discussion (10 minutes)	The class compares predictions with what was observed; the teacher names the idea behind the wrong predictions	The class discusses the observations; the teacher draws out the concept
Closing (5 minutes)	Summary and exit question	Summary and exit question

Note. The two lessons used the same competencies, worked examples, and homework. POE = Predict-Observe-Explain.

The Simulations and the Tasks

Each unit had four POE tasks, one every two lessons. Unit 1 used two screens of PhET's Forces and Motion: Basics (Version 2.6.7) for three tasks: the Motion screen, where the crate rides a frictionless skateboard, to watch it

after the push ends; and the Acceleration screen, to apply a steady force and to double the mass. The fourth task, on action-reaction pairs, used two spring balances hooked together, because the simulation does not show the pair of forces. Unit 2 used PhET's Circuit Construction Kit: DC (Version 1.5.4): ammeters before and after a bulb, a resistor added after a bulb, and bulbs added in parallel. The simulations were downloaded as HTML5 files and run offline in a web browser on the 20 laptops of the school's computer laboratory, so no lesson depended on the internet. PhET's simulations are free for classroom use under a Creative Commons Attribution-NonCommercial 4.0 license, and each worksheet carried PhET's attribution.

Strategy and Evaluation Plan

Every learner received both methods. Each section had POE lessons in one unit and the regular lessons in the other, in opposite orders, so that neither section was denied the innovation and each learner's two units could be compared. The success criteria, set before the units began, were a posttest advantage for the POE unit that was statistically significant at the .05 level after accounting for the unit and the order, and fewer misconception answers on the posttests of the POE units. The co-researcher, a Master Teacher in the school's science department, observed twelve lessons with a fidelity checklist (Appendix A).

Action Research Questions

The study answered five questions:

1. What were the learners' pretest and posttest scores in each unit under the POE lessons with simulations and under the regular lessons?

2. Is there a significant difference between the learners' posttest scores under the two methods when the unit, the order of the methods, and the pretest are taken into account?
3. Is there a significant difference between the number of misconception answers the learners chose on the posttests under the two methods?
4. Which misconceptions did the learners show in their predictions, and which remained after the units?
5. How did the learners describe learning with POE tasks and simulations?

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

H_{01}: There is no significant difference between the learners' posttest scores under the POE lessons with simulations and under the regular lessons, taking the unit, the order of the methods, and the pretest into account.

H_{02}: There is no significant difference between the number of misconception answers the learners chose on the posttests under the POE lessons with simulations and under the regular lessons.

Action Research Methods

Research Design

The study used a two-period crossover design, AB/BA (Jones & Kenward, 2014; Senn, 2002). The two periods were the first two units of Grade 9 Science in Term 1 of School Year 2026–2027, set by the budget of work for the grade (DepEd, 2026c): Unit 1, Newton's laws of motion, in weeks 1 and 2 (June 15 to 26, 2026), and Unit 2, electric current and circuits, the first three competencies of weeks 3 to 8, taught in weeks 3 and 4 (June 29 to July 10).

The two sections were intact classes. A coin toss on June 11, during the Opening Block, decided that Narra would have POE lessons first.

Table 2

The Crossover Design

Section	Unit 1, Newton's laws (June 15–26)	Unit 2, electric circuits (June 29–July 10)
Narra (*n* = 41)	POE lessons with simulations	Regular lessons
Molave (*n* = 39)	Regular lessons	POE lessons with simulations

Note. Each unit began with its pretest and ended with its posttest. Unit 1 had eight lessons and Unit 2 seven: July 6, the day of Term 1's first teacher-made summative test, fell in Unit 2. POE = Predict-Observe-Explain.

The design compares each learner with himself or herself, which removes the large differences between learners from the comparison of methods, and it compares the sections within each unit, which removes the difference between the units. What it cannot separate is the order of the methods from the difference between the sections: any carryover of POE habits from Narra's first unit into its second is mixed with Narra's own strengths (Freeman, 1989). The two units had different content, which limits the carryover of knowledge, and the pretests were used to check the sections' starting points.

Participants and Other Sources of Data and Information

The participants were the 80 learners of the two Grade 9 sections taught Science by the lead researcher.

Table 3

Learners by Section and Sex

Section	Female	Male	Total
Narra	27	14	41
Molave	21	18	39

Section	Female	Male	Total
Total	48	32	80

Note. Two Molave learners transferred to another school in week 3 and have Unit 1 data only; 78 learners took both units.

Before the study, a power analysis was run by simulation in Python. With 40 learners in each section, a posttest standard deviation of 16 points, a correlation of .50 between a learner's two posttests, Unit 2 eight points harder, and $\alpha = .05$, 4,000 simulated studies gave a power of .80 to detect a difference of 5 percentage points between the methods, and .91 for 6 points.

Data Gathering Methods

Unit Tests

Each unit had a 25-item multiple-choice test with four options, written by the researchers from the unit's competencies (Appendix A). Every wrong option was written to catch either one of eight misconceptions from the research (Hestenes et al., 1992; McDermott & Shaffer, 1992; Shipstone, 1984) or none: four on force and motion and four on circuits (Appendix C gives one item for each). Thirty items were drafted for each unit. Five experts, two Master Teachers in Science, a Head Teacher, the division's Education Program Supervisor for Science, and a physics professor at a state university, rated each draft as essential, useful but not essential, or not necessary. Lawshe's (1975) content validity ratio was computed for each; with five experts, the minimum is .99, so a draft stayed only when all five rated it essential, and four drafts were dropped.

The remaining 56 drafts were piloted on June 9 and 10, during the Opening Block, with 40 Grade 10 learners who would not take part in the study. For each draft, the difficulty index p was the share who answered it

correctly, and the discrimination index *D* was the share correct in the upper 27% of the unit's total score minus the share correct in the lower 27% (Kelley, 1939), 11 learners in each group. A draft was kept when .20 ≤ *p* ≤ .85 and *D* ≥ .20; Ebel and Frisbie (1991) regard a *D* of .20 to .29 as marginal and of .40 or more as very good. A distractor was counted as functional when at least 5% of learners chose it and the lower group chose it more often than the upper group (Haladyna & Downing, 1993; Tarrant et al., 2009).

Table 4

Item Analysis of the Pilot

Statistic	Unit 1	Unit 2
Drafts after the expert panel	28	28
Drafts dropped after the pilot	3	3
Items on the final test	25	25
Mean *p* of the final items	.59	.51
Range of *p*	.38–.83	.33–.80
Mean *D* of the final items	.54	.51
Nonfunctional distractors, of 75	22	20
KR-20, drafts	.80	.78
KR-20, final test	.84	.81

Note. Pilot with 40 Grade 10 learners, June 9 and 10, 2026. The six dropped drafts were two that nearly everyone answered correctly (*p* = .93 and 1.00) and four whose *D* was zero or negative. KR-20 = Kuder-Richardson Formula 20.

The tests' reliability was estimated with the Kuder-Richardson Formula 20 (Kuder & Richardson, 1937). The 42 nonfunctional distractors, 10 of which carried a misconception the study counted, were kept so that the tests stayed as piloted, and were noted for revision.

Each test was given as a pretest on the first day of its unit and as a posttest on the last, in 30 minutes. Answers were not discussed and scores were not returned until both posttests were done, and the scores did not count toward the learners' grades.

POE Worksheets

On each POE worksheet, the learner wrote a prediction and its reason before running the simulation. The researchers sorted each prediction as correct, as showing the misconception the task was written to reveal, or as other, and a second researcher checked a random fifth of them.

Fidelity Checklists

The co-researcher observed six lessons in each unit, three in each section, with a five-step checklist for each method (Appendix A).

Focus Groups

On July 13 and 14, after both posttests, one focus group of six learners from each section discussed learning with the POE tasks and the simulations for about 40 minutes after classes (Appendix D). The learners were chosen to include two with high, two with middle, and two with low Unit 1 posttest scores, and they spoke in English or Filipino as they chose.

Ethical Considerations

The Schools Division Research Committee approved the proposal before the study began; its reference number is withheld for blind review. The school head gave written permission, parents gave written consent at the June 9 orientation, and learners gave their assent; the Grade 10 pilot learners' parents gave their consent, and the learners their assent, before the pilot. Because the design gave both methods to every learner, no section was denied the innovation. The unit tests did not count toward grades, and learners were told that their answers on the worksheets and in the focus groups would not affect their grades. Learners were identified only by codes, the focus group

recordings were deleted after transcription, and the data were kept on a password-protected computer, one of the security measures the Data Privacy Act of 2012 requires for personal information (Republic Act No. 10173, 2012, Sec. 20). The researchers declared their use of an artificial intelligence (AI) tool in Appendix E (DepEd, 2026a).

Data Analysis

Scores were converted to percentages of the 25 items. Means and standard deviations described the scores by unit and method. Welch's t test compared the two sections' pretests in each unit.

The posttest scores were analyzed with a linear mixed-effects model fitted by restricted maximum likelihood in the statsmodels library for Python (Seabold & Perktold, 2010): fixed effects for the method (POE or regular), the unit, the section, and the unit's pretest score centered on its mean, and a random intercept for each learner, which allows for the correlation between a learner's two scores. The model keeps the two learners who left after Unit 1. The method's coefficient, with its 95% confidence interval, answered the second question. Three checks followed: the model without the pretest; the model with the 78 learners who took both units; and the Hills and Armitage (1979) test, a two-sample t test on each learner's Unit 1 score minus Unit 2 score, which is the classical analysis of a crossover, with its nonparametric counterpart, the Mann-Whitney test on the same differences (Koch, 1972). The same test was run as a mixed analysis of variance in jamovi, where the unit × section interaction is the method's effect (The jamovi project, 2026).

The residuals were checked for normality with the Shapiro-Wilk test and a quantile plot.

The effect size was Cohen's *d*: the Hills and Armitage estimate of the method's effect divided by the pooled standard deviation of the posttest scores in the four unit-method cells (Lakens, 2013). Its 95% confidence interval came from 2,000 cluster bootstrap samples (Efron & Tibshirani, 1993), which resampled whole learners within each section, so that each learner's two scores stayed together.

For the third question, the number of a learner's 25 posttest answers that chose an option carrying a misconception was analyzed with generalized estimating equations (GEE) with a Poisson distribution, an exchangeable working correlation within each learner, and robust standard errors (Liang & Zeger, 1986), with the method, the unit, the section, and the pretest count as predictors. The exponentiated coefficient is a rate ratio. Holm's (1979) procedure adjusted the *p* values of the two hypotheses for testing two. The predictions were summarized as percentages, and the focus groups were analyzed thematically in Braun and Clarke's (2006) six phases.

All analyses were run in Python 3.12 with pandas, SciPy, statsmodels, and pingouin (Vallat, 2018), free software that a school can use without buying a license; the notebook and the Excel data set reproduce every result (Appendix A).

Discussion of Results and Reflection

Implementation

The co-researcher observed 58 of the 60 checklist steps in the twelve lessons (96.7%). In two POE lessons, on June 19 in Narra and July 8 in Molave, the written explanation was cut short by the bell and finished as homework; every other POE step and every step of the regular lessons was observed. The groups shared the 20 laptops, two or three learners to a laptop, and no lesson lost time to the internet. Two Molave learners transferred to another school in week 3; their Unit 1 scores are kept in the mixed model.

Question 1: Pretest and Posttest Scores

The two sections started even. On the Unit 1 pretest, Narra averaged 29.9% and Molave 32.2%, Welch's $t(78.0) = -0.76$, $p = .449$. Both pretests were near the 25% that guessing among four options would give, because most learners chose the options carrying everyday ideas.

Table 5

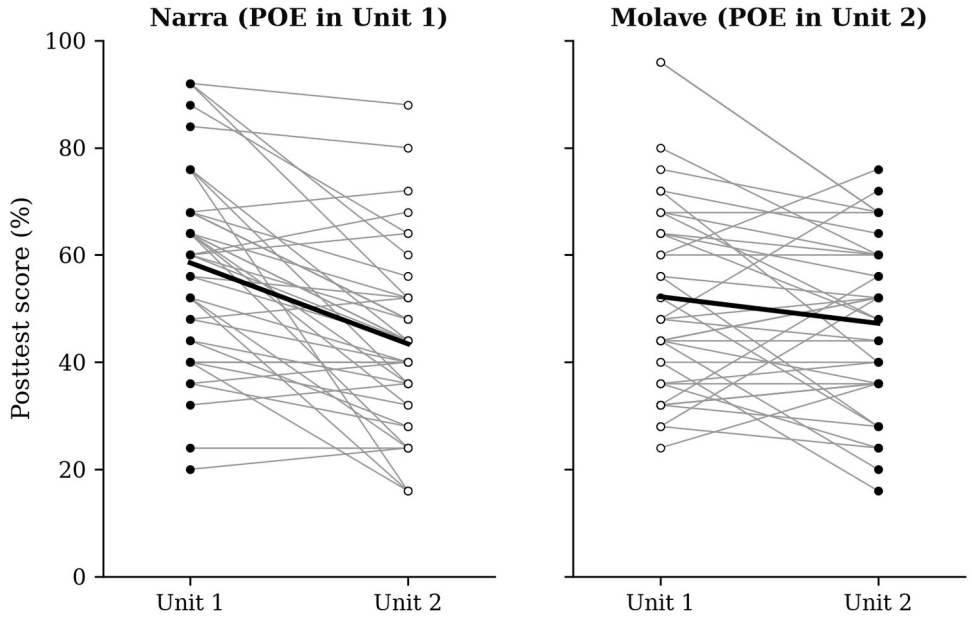
Pretest and Posttest Scores (%) by Unit and Method

Unit	Method	Section	*n*	Pretest *M*	Pretest *SD*	Posttest *M*	Posttest *SD*
1. Newton's laws	POE	Narra	41	29.9	14.0	58.5	18.0
1. Newton's laws	Regular	Molave	39	32.2	13.6	52.2	18.6
2. Circuits	POE	Molave	37	26.3	10.0	47.2	15.9
2. Circuits	Regular	Narra	41	25.8	11.4	43.4	16.7

Note. Scores are percentages of 25 items. POE = Predict-Observe-Explain lessons with simulations.

Figure 1

Each Learner's Posttest Scores in the Two Units, by Section

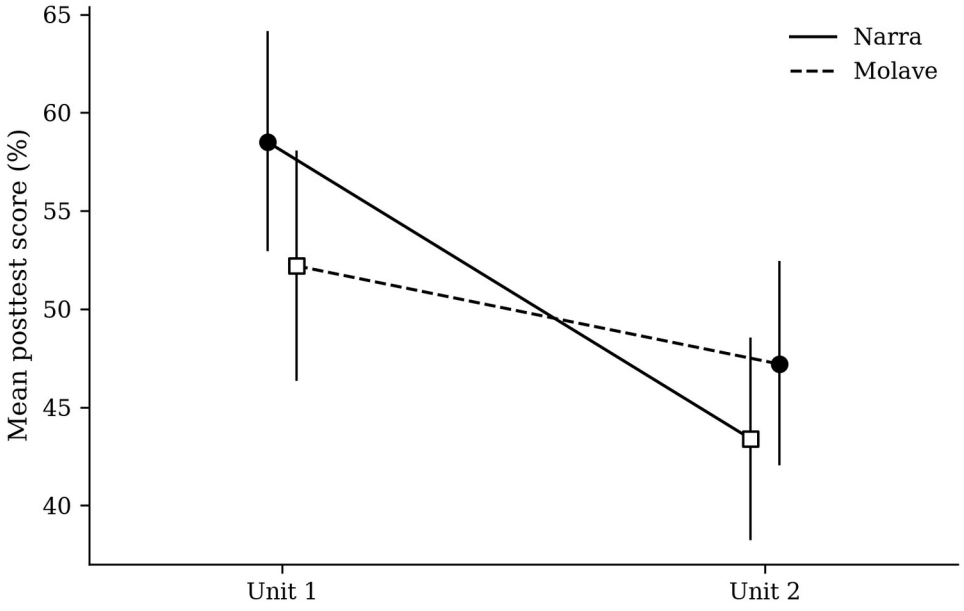


Note. Each line joins one learner's two posttest scores; filled markers show the unit with POE lessons. The heavy line joins the section's means. The two Molave learners who left after Unit 1 have one score each.

In each unit, the section with POE lessons had the higher posttest mean (Table 5). Unit 2 was harder for both: Narra's mean fell by 15.1 points from Unit 1 to Unit 2 and Molave's by 5.0 (Figure 1), and the difference between those two falls is the evidence of the method's effect.

Figure 2

Mean Posttest Scores by Unit and Section



Note. Error bars are 95% confidence intervals. Filled markers show POE lessons, open markers regular lessons.

The reliability of the tests in the study's sample was lower than in the pilot. The posttests' KR-20 was .75 for Unit 1 and .67 for Unit 2, and the pretests', taken near the level of guessing, .58 and .35.

Question 2: The Difference in Posttest Scores

Table 6

Linear Mixed-Effects Model of the Posttest Score (%)

Term	*b*	*SE*	*z*	*p*	95% CI
Intercept	51.46	2.35	21.89	< .001	[46.85, 56.06]
POE lessons (vs. regular)	5.68	1.85	3.06	.002	[2.05, 9.31]
Unit 2 (vs. Unit 1)	−10.01	1.85	−5.41	< .001	[−13.63, −6.38]
Narra (vs. Molave)	2.12	2.73	0.77	.438	[−3.24, 7.48]
Pretest, centered (per point)	0.62	0.11	5.63	< .001	[0.41, 0.84]

Note. 158 posttest scores of 80 learners. Variance between learners = 80.6; residual variance = 134.2; intraclass correlation = .38. Restricted maximum likelihood (statsmodels MixedLM). POE = Predict-Observe-Explain.

Taking the unit, the section, and the pretest into account, learners scored 5.68 percentage points higher on the posttest of their POE unit, 95% CI [2.05, 9.31], $z = 3.06$, $p = .002$ (Table 6); with Holm's adjustment for two hypotheses, $p = .002$. The first null hypothesis is rejected. Unit 2 was harder by 10.01 points, and the sections did not differ once the units and pretests were accounted for.

Table 7

The POE Effect (Percentage Points) by Analysis

Analysis	Estimate	95% CI	Test	*p*
Mixed model, with pretest (main)	5.68	[2.05, 9.31]	$z = 3.06$	.002
Mixed model, without pretest	5.08	[1.67, 8.48]	$z = 2.92$	.003
Mixed model, 78 learners with both units	5.70	[2.07, 9.33]	$z = 3.07$	.002
Hills and Armitage *t* test on the period differences	5.07	[1.61, 8.54]	$t(76) = 2.91$	.005
Mixed ANOVA, unit × section interaction			$F(1, 76) = 8.50$	.005
Mann-Whitney test on the period differences			$U = 508$	.012

Note. The mixed ANOVA's F is the square of the Hills and Armitage t; partial $\eta^2 = .101$. POE = Predict-Observe-Explain.

The checks agree (Table 7). Without the pretest, the estimate is 5.08 points, and the classical crossover test gives 5.07 points, $t(76) = 2.91$, $p = .005$. The effect size is small: $d = 0.29$, 95% bootstrap CI [0.10, 0.51]. The model's residuals were close to normal, Shapiro-Wilk $W = .997$, $p = .983$, with similar spread in the four unit-method cells (standard deviations 8.6 to 10.7).

A paired t test of each learner's POE score against the regular score would have given a mean difference of 5.59 points, $t(77) = 2.70$, $p = .009$, but its differences mix the method with the unit: they averaged 15.12 points in Narra, whose POE unit was the easier one, and −4.97 in Molave. Their

standard deviation, 18.29, was larger than either section's own (15.66 and 15.01), so the test's interval was wider than the crossover's, and a larger difference in the sections' sizes would have biased its mean toward whichever section was larger. The crossover analyses keep the unit out of the comparison.

Question 3: Misconception Answers

On the posttests, learners chose fewer options carrying a misconception after POE lessons: 7.98 of 25 answers in Unit 1 against 9.05 under the regular lessons, and 10.46 in Unit 2 against 11.90. Taking the unit, the section, and the pretest count into account, the rate of misconception answers under POE lessons was 0.86 times the rate under the regular lessons, 95% CI [0.79, 0.92], $z = -4.11$, $p < .001$; Holm-adjusted $p < .001$. The second null hypothesis is rejected. The Pearson dispersion was 1.19, close to the 1 a Poisson model expects, and the robust standard errors allow for the rest.

Question 4: Misconceptions in the Predictions and the Posttests

The POE worksheets show the misconceptions before any observation. On each of the seven tasks written to reveal a misconception, more than half of the learners predicted what the misconception would predict (Table 8). The most common was the idea that a change in a circuit affects only what comes after it: 73.0% of Molave's learners predicted that a resistor placed after a bulb would not change the bulb.

Table 8

Predictions on the POE Worksheets

Task	What learners predicted	Misconception	*n*	Correct (%)	Misconception (%)	Other (%)
1A	What a crate does after the push ends, with no friction	F1	41	34.1	51.2	14.6
1B	The crate's speed under a steady net force	F2	41	34.1	51.2	14.6
1C	The acceleration when a second crate is stacked on the first	None	41	51.2		48.8
1D	The readings of two hooked spring balances when one learner pulls hard	F3	41	17.1	63.4	19.5
2A	Ammeter readings before and after a bulb	C1	37	21.6	67.6	10.8
2B	A bulb when a resistor is added after it	C3	37	18.9	73.0	8.1
2C	The first bulb and the battery's current when a bulb is added in parallel	C2	37	27.0	51.4	21.6
2D	The total resistance when a third bulb is added in parallel	C4	37	24.3	59.5	16.2

Note. Narra did tasks 1A to 1D in Unit 1 and Molave tasks 2A to 2D in Unit 2. F1 = motion needs a force; F2 = constant force, constant speed; F3 = the stronger pusher pushes harder; C1 = current is used up; C2 = the battery gives a fixed current; C3 = only what comes after changes; C4 = more bulbs, more resistance.

After teaching, every misconception was chosen less than on the pretest under both methods, and seven of the eight were chosen less after POE lessons than after the regular lessons (Table 9, Figure 3). The largest differences were for the idea that an object at rest has no forces on it (29.3% against 40.4%) and for the two ideas about current and the battery (C1 and C2, about 8 points each). The idea that more bulbs in parallel means more resistance was the exception: 39.6% after POE lessons against 36.6% after the regular lessons.

Table 9

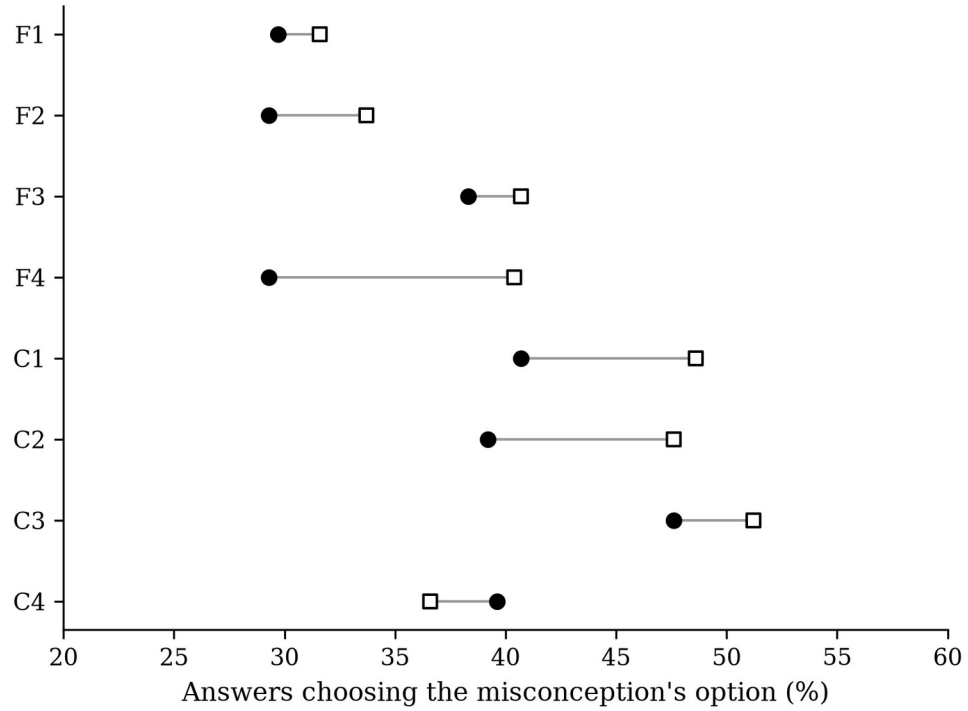
Answers Choosing a Misconception's Option (%), on the Items Carrying It

Code	Misconception	Items	Pretest, POE	Pretest, regular	Posttest, POE	Posttest, regular
F1	Motion needs a force	6	56.9	59.0	29.7	31.6
F2	Constant force, constant speed	8	56.7	50.3	29.3	33.7
F3	The stronger pusher pushes harder	7	65.5	55.7	38.3	40.7
F4	At rest, no forces	4	59.1	59.0	29.3	40.4
C1	Current is used up	15	62.0	62.8	40.7	48.6
C2	The battery gives a fixed current	2	60.8	59.8	39.2	47.6
C3	Only what comes after changes	5	64.9	66.8	47.6	51.2
C4	More bulbs, more resistance	3	60.4	56.1	39.6	36.6

Note. Each percentage is the share of answers, on the items with an option carrying the misconception, that chose such an option. Unit 1 (F1 to F4): POE is Narra, regular is Molave; Unit 2 (C1 to C4): POE is Molave, regular is Narra.

Figure 3

Posttest Answers Choosing a Misconception's Option, by Method



Note. Filled circles show the POE unit; open squares the regular unit.

Question 5: The Learners' Accounts

Four themes came from the two focus groups (Table 10).

Table 10

Themes From the Focus Groups

Theme	Learners (of 12)	A learner's words
A wrong prediction is hard to forget	10	"Akala ko titigil yung kahon, pero tuloy-tuloy lang pala. Doon ko na-gets." ("I thought the box would stop, but it just kept going. That's when I got it.")
Seeing what cannot be seen	9	"In the simulation you can see the current going around, the same before and after the bulb."
The simulation is not the real thing	7	"Sa totoo, humihina yung baterya namin sa bahay. Sa computer, hindi." ("In real life, our battery at home gets weak. On the computer, it doesn't.")
Sharing a laptop	6	"When three of us share, one clicks and two only watch."

Note. Six learners from each section, July 13 and 14, 2026. A learner may voice more than one theme. Translations are the researchers'.

The first theme matches the purpose of the prediction step: learners remembered most where their prediction failed. The third is a caution. Learners noticed that the simulation's battery does not weaken and that its crate never meets friction, and some doubted that the rules hold outside the screen; the teacher had to link each simulation back to a real cart and a real battery. The fourth points to the number of laptops.

Reflection

POE lessons with simulations were followed by higher posttest scores and fewer misconception answers in both units, in both orders, and the result held across the checks. The effect was small, $d = 0.29$; in points, it was about half the 10-point difference between the two units themselves. It is in line with the evidence that simulations help most when they replace or support

laboratory work (Finkelstein et al., 2005; Rutten et al., 2012) and with the idea that a written prediction makes a learner's own belief something an observation can test (White & Gunstone, 1992). The predictions themselves were useful: they showed the teacher, lesson by lesson, which ideas the class held, before any test.

The study has limits. The innovation joined two things, the POE sequence and the simulation, so it cannot say which helped. The two sections were intact classes, and the design cannot separate the order of the methods from the sections' own differences; the Unit 2 pretests were nearly equal, 25.8% in Narra and 26.3% in Molave, Welch's $t(75.9) = -0.21$, $p = .832$, so Narra did not start Unit 2 knowing more, but carryover in how learners study cannot be ruled out. The lead researcher taught both methods and knew which was being tested. The Unit 2 posttest's KR-20 in the study's sample, .67, was below the .81 of the pilot and the .70 the researchers had aimed for: reliability belongs to a set of scores, not to a test, and the study's learners, who all came from one grade, varied less than the pilot's. Lower reliability adds noise to the scores and makes a difference harder to detect. One misconception, more bulbs meaning more resistance, did not fall more under POE lessons, perhaps because the two parallel-circuit tasks came last in Unit 2 and one of them was cut short by the bell. Finally, the study lasted four weeks; it cannot show whether the advantage lasts to the end of the term.

Action Plan

Table 11

Action Plan

Activity	Objective	Persons involved	Timeline	Resources	Success indicator
Present the results at a learning action cell (LAC) session of the science department	Decide which Science units use POE lessons with simulations	Researchers, science teachers, school head	August 2026	Report, one-page brief	Decision recorded in the LAC minutes
Use POE tasks in the units where learners' everyday ideas are common, such as force and motion, circuits, and chemical bonding in Term 3	Bring learners' predictions into the open and test them	Grade 9 and Grade 10 science teachers	Terms 2 and 3, SY 2026–2027	Offline simulations, worksheets	POE worksheets on file for each unit
Revise the parallel-circuit task and the 22 and 20 nonfunctional distractors	Strengthen the weakest task and the tests	Researchers	September 2026	Item analysis	Revised tests piloted
Ask the school governing council and the parent-teacher association for ten more laptops for the computer laboratory	Reduce the groups to two learners a laptop	School head, ICT coordinator	December 2026	School and PTA funds	Laptops in use
Link each simulation to a real setup in the same lesson	Answer learners' doubt that the rules hold off the screen	Science teachers	From Term 2	Laboratory materials	Real setup listed in each POE lesson plan
Propose a follow-up study with a delayed posttest at the end of the term	Test whether the advantage lasts	Researchers	2027 BERF call	BERF grant	Proposal submitted

References

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Department of Education. (2016). *Adoption of the Basic Education Research Agenda* (DepEd Order No. 39, s. 2016).

Department of Education. (2017). *Research management guidelines* (DepEd Order No. 16, s. 2017).

Department of Education. (2021). *Amendment to DepEd Order No. 016, s. 2017 (Research management guidelines)* (DepEd Order No. 026, s. 2021).

Department of Education. (2026a). *Foundational guidelines on artificial intelligence (AI) in basic education* (DepEd Order No. 003, s. 2026).

Department of Education. (2026b). *Guidelines on the implementation of the three-term school calendar in basic education* (DepEd Order No. 009, s. 2026).

Department of Education. (2026c). *Grade 9 Science* [Budget of work, School Year 2026–2027].

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice Hall.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall.

Finkelstein, N. D., Adams, W. K., Keller, C. J., Kohl, P. B., Perkins, K. K., Podolefsky, N. S., Reid, S., & LeMaster, R. (2005). When learning

about the real world is better done virtually: A study of substituting computer simulations for laboratory equipment. *Physical Review Special Topics—Physics Education Research, 1*(1), Article 010103. https://doi.org/10.1103/PhysRevSTPER.1.010103

Freeman, P. R. (1989). The performance of the two-stage analysis of two-treatment, two-period crossover trials. *Statistics in Medicine, 8*(12), 1421–1432. https://doi.org/10.1002/sim.4780081202

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement, 53*(4), 999–1010. https://doi.org/10.1177/0013164493053004013

Halloun, I. A., & Hestenes, D. (1985). Common sense concepts about motion. *American Journal of Physics, 53*(11), 1056–1065. https://doi.org/10.1119/1.14031

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30*(3), 141–158. https://doi.org/10.1119/1.2343497

Hills, M., & Armitage, P. (1979). The two-period cross-over clinical trial. *British Journal of Clinical Pharmacology, 8*(1), 7–20. https://doi.org/10.1111/j.1365-2125.1979.tb05903.x

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

Jones, B., & Kenward, M. G. (2014). *Design and analysis of cross-over trials* (3rd ed.). CRC Press.

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30*(1), 17–24. https://doi.org/10.1037/h0057123

Koch, G. G. (1972). The use of non-parametric methods in the statistical analysis of the two-period change-over design. *Biometrics, 28*(2), 577–584. https://doi.org/10.2307/2556170

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151–160. https://doi.org/10.1007/BF02288391

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology, 4*, Article 863. https://doi.org/10.3389/fpsyg.2013.00863

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

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*(1), 13–22. https://doi.org/10.1093/biomet/73.1.13

McDermott, L. C., & Shaffer, P. S. (1992). Research as a guide for curriculum development: An example from introductory electricity. Part I: Investigation of student understanding. *American Journal of Physics, 60*(11), 994–1003. https://doi.org/10.1119/1.17003

PhET Interactive Simulations. (n.d.-a). *Circuit Construction Kit: DC* (Version 1.5.4) [Computer simulation]. University of Colorado Boulder. Retrieved May 18, 2026, from https://phet.colorado.edu/en/simulations/circuit-construction-kit-dc

PhET Interactive Simulations. (n.d.-b). *Forces and Motion: Basics* (Version 2.6.7) [Computer simulation]. University of Colorado Boulder. Retrieved May 18, 2026, from https://phet.colorado.edu/en/simulations/forces-and-motion-basics

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

Rutten, N., van Joolingen, W. R., & van der Veen, J. T. (2012). The learning effects of computer simulations in science education. *Computers & Education, 58*(1), 136–153. https://doi.org/10.1016/j.compedu.2011.07.017

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

Senn, S. (2002). *Cross-over trials in clinical research* (2nd ed.). Wiley.

Shipstone, D. M. (1984). A study of children's understanding of electricity in simple DC circuits. *European Journal of Science Education, 6*(2), 185–198. https://doi.org/10.1080/0140528840060208

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice

questions: A descriptive analysis. *BMC Medical Education, 9*, Article 40. https://doi.org/10.1186/1472-6920-9-40

The jamovi project. (2026). *jamovi* (Version 2.7) [Computer software]. https://www.jamovi.org

Vallat, R. (2018). Pingouin: Statistics in Python. *Journal of Open Source Software, 3*(31), Article 1026. https://doi.org/10.21105/joss.01026

White, R., & Gunstone, R. (1992). *Probing understanding*. Falmer Press.

Wieman, C. E., Adams, W. K., & Perkins, K. K. (2008). PhET: Simulations that enhance learning. *Science, 322*(5902), 682–683. https://doi.org/10.1126/science.1161948

Financial Report

The study was funded by a BERF grant of PhP 7,500.00 for school-based action research. Under DepEd Order No. 16, s. 2017, 80% (PhP 6,000.00) was released on submission of the work plan and the signed memorandum of agreement, and the last 20% (PhP 1,500.00) is released when the research committee accepts the final report (DepEd, 2017). The simulations, Python, and jamovi cost nothing, and the laptops were the school's; the grant paid for no software, equipment, or statistician, which BERF does not cover.

Table 12

Statement of Expenses

Date (2026)	Particulars	Receipt No.	Category	Budget (PhP)	Actual (PhP)
May 29	Photocopying of the draft tests and rating sheets for the five experts (300 pages)	OR 0511	Reproduction	300.00	300.00
June 1	Travel to deliver and collect the experts' rating sheets	Acknowledgment receipt	Domestic travel	400.00	400.00
June 5	Consent and assent forms and letters to parents (300 pages)	OR 0538	Reproduction	300.00	300.00
June 5	Pilot test booklets and answer sheets (400 pages)	OR 0539	Reproduction	400.00	400.00
June 11	Unit test booklets and answer sheets, pretests and posttests (1,600 pages)	OR 0561	Reproduction	1,600.00	1,560.00
June 11	POE worksheets and regular activity sheets (1,000 pages)	OR 0562	Reproduction	1,000.00	1,040.00
June and July	Prepaid load for coordination with the experts and parents, two months	OR 3317, 3359	Communication	600.00	600.00
July 13 and 14	Snacks for the two focus groups (12 learners and 2 researchers)	OR 0712, 0713	Food during data gathering	900.00	900.00

Date (2026)	Particulars	Receipt No.	Category	Budget (PhP)	Actual (PhP)
September 21	Printing and binding of the final report (4 copies)	OR 0904	Printing and binding	1,400.00	1,400.00
September 22	One-page brief for the science department and parents (300 copies)	OR 0905	Reproduction	600.00	600.00
	Total			7,500.00	7,500.00

Note. Receipt numbers are invented for this sample.

Appendices

Appendix A. Data Set, Notebook, and Instruments

The executive data set is the Excel workbook submitted with this report: every learner's pretest and posttest scores, every answer with its scoring and misconception code as formulas, the 60 draft items with the experts' ratings and Lawshe's CVR, the pilot's answers with the item analysis and KR-20 as formulas, the crossover test on the period differences, the predictions, and the fidelity checklists. The mixed model, the GEE, the bootstrap, and the power simulation come from the Jupyter notebook submitted with it, which reads the workbook and reproduces every result. The fidelity checklist for POE lessons had five steps: prediction written before the simulation; reason written with the prediction; simulation run by the groups; observation compared with the prediction; explanation revised in writing. The checklist for regular lessons: the lesson exemplar's activity followed; no simulation used; same competency and time as the POE lesson; observations recorded by the learners; concept summarized by the teacher.

Appendix B. Letters, Consent, and Assent Forms

The school head's permission, the letter to parents, and the consent and assent forms are kept at the school and left out of the archived report.

Appendix C. Sample Items

One item for each misconception, with its key and the code of each option, is printed in the data set's Sample items sheet; the other items are kept at the school, so the tests can be used again.

Appendix D. Focus Group Guide

1. What did you do in a lesson with the simulation?
2. Tell us about a time your prediction was wrong. What happened next?
3. What did the simulation show you that the regular activity did not, or the other way around?
4. When does the simulation not match what you see in real life?
5. What would make these lessons work better for you?

Appendix E. Declaration of AI Use

We, the researchers, declare that we used [name of the AI tool] only to check the grammar of the English text of this report, for the purpose of finalizing the research study. The prompt used was: "Check the grammar of this paragraph. Do not change its meaning or add content." We have reviewed, edited, and understood all the content, and we take full accountability for it. No names, codes, scores, worksheets, recordings, or other data from the study were entered into the AI tool.