

Magna Education and AP Exam Outcomes: Study Report

Oct 27, 2026 · Magna Education

In a quasi-experimental historical-cohort study with student-level adjustment for pre-intervention achievement, AP U.S. Government students in the Magna cohort had 2.25 times the odds of scoring 4 or 5 on the College Board AP exam compared with the same teacher's prior-year cohort (95% CI 1.04 to 4.87, $p = .039$; Hedges' $g = 0.49$). Their adjusted mean AP score was 0.32 points higher (0.03 to 0.61, $p = .032$; $g = 0.33$), a result confirmed by exact baseline matching and a permutation test. A supporting class-level comparison in AP Biology, with equivalent pre-intervention scores, showed a pass-rate gain of 10.1 percentage points that was not statistically significant; its share scoring 4 or 5 did not rise. The findings are consistent with the requirements for ESSA Tier 3 (Promising Evidence).

Results at a glance

In both studies the comparison group is the same teacher's class from the year before Magna, and the outcome is the independently scored College Board AP exam. National results are reported as context, not as a control group.

	Study 1: AP U.S. Government (qualifying study)	Study 2: AP Biology (supporting replication)
Design	Quasi-experimental historical-cohort study with a nonequivalent comparison group and student-level statistical adjustment for pre-intervention achievement	Historical-cohort comparison with equivalent class-level pre-intervention achievement, adjusted for the national change
Exam years	2024 (comparison) vs 2025 (Magna)	2025 (comparison) vs 2026 (Magna)
Students	79 vs 60 (139)	About 88 vs 87
Pre-intervention measure	Unit 1 test, student level	Unit 1 test, class mean: 76.9% vs 76.7%

	Study 1: AP U.S. Government (qualifying study)	Study 2: AP Biology (supporting replication)
Primary outcome	Scored 4 or higher: odds ratio 2.25 (1.04 to 4.87), $p = .039$; Hedges' $g = 0.49$	Pass rate (3 or higher): 78.4% to 88.5%, +9.3 points net of the national change, $p = .09$
Secondary outcomes	Scored 5: odds ratio 3.25 (1.25 to 8.49), $p = .016$. Mean AP score: +0.32 (0.03 to 0.61), $p = .032$; $g = 0.33$. Scored 3 or higher: odds ratio 0.91 (0.29 to 2.83), $p = .86$	Scored 1 or 2: 21.6% to 11.5%, $p = .10$. Scored 4 or 5: 54% to 50%, about -1 point net of the national change, $p = .65$. Mean AP score: +0.10 net of the national change, $p = .53$
Follow-up	2026 cohort ($n = 78$): 59% scored 4 or higher vs 51% nationally (descriptive)	—

Introduction

AP exam results matter to students and schools. Scores of 3 or higher can earn college credit, and scores of 4 and 5 qualify for credit or placement at more colleges. The free-response section is where many students lose the most points.

The problem targeted. Nationally, AP students earn far fewer free-response points when they must explain than when they must identify. On the 2025 AP U.S. Government and Politics Quantitative Analysis question, students earned 94% of the “identify” and “describe” points but 39% of the “draw a conclusion” and “explain” points, a simple average of the two exam sets’ part-level means ([2025 Chief Reader Report, Set 1, Set 2](#)). Chief Readers describe the same pattern in both study courses: AP Gov responses that “simply identified the powers without demonstrating understanding of the application of these powers” ([2024 Chief Reader Report, Set 1](#)), and AP Biology responses showing “the inability to justify a claim” ([2025 Chief Reader Report](#)). The Study 1 class showed the same gap locally in 2024–25: 71.6% on multiple-choice practice versus 36.5% on free response.

About Magna Education

Magna Education AI is an education technology company based in San Francisco, California. Its platform gives students AI-scored practice and feedback in 13 AP courses and the TSIA2 college-readiness test, and gives teachers class- and student-level reports.

Founder and CEO Sarvesh Bhardwaj holds a PhD in electrical engineering from Arizona State University and an MBA from the Wharton School, and spent more than 15 years at Amazon Web Services, Amazon Lab126 and Mentor Graphics. Magna received a 2025 Top EdTech Products of the Year award from *District Administration* magazine.

Evaluation questions

1. After accounting for prior achievement, did students in a classroom using Magna score higher on the AP exam than the same teacher's previous class? (Study 1; primary outcome: scoring 4 or higher)
2. Did the free-response question types students practiced in Magna improve more than the type they did not practice? (Study 1, exploratory)
3. Did a second course, taught by a second teacher, show similar results? (Study 2)
4. How did teachers and students describe their use of Magna? (Educator and student voices)

Key design features

Five features of the design shape how these results should be read.

- **Comparison group.** In both studies the comparison group is the same teacher's class from the year before Magna, not a class in the same year.
- **How the classrooms were chosen.** They are the only Magna classrooms with a pass rate below 85% in the year before Magna and a comparable unit test given before Magna use. Selecting classrooms on a below-threshold prior year can by itself make the next year look better (regression to the mean).
- **Pass rate.** In Study 1 the share scoring 3 or higher did not change significantly after adjustment (odds ratio 0.91, $p = .86$); the gains were in scores of 4 and 5.
- **Exam change.** The AP U.S. Government exam moved to a fully digital format in May 2025, the Magna year.
- **Who analyzed the data.** The Study 1 teacher collected the data. Magna Education analyzed it retrospectively, without a pre-registered primary outcome.

Program description and logic model

Magna Education is an AI-powered practice and feedback platform for AP courses and TSIA2, used alongside regular instruction. It generates AP-style multiple-choice and free-response practice aligned to the College Board course framework. It scores free-response answers against College Board rubrics and returns feedback the moment a student submits. Students then use Ask Magna, an AI tutor that is grounded in their own submitted work, stays within course content, and asks guiding questions before giving an answer, to clear up misconceptions. Skill Builder, introduced after both studies, breaks each question into its problem-solving and critical-thinking sub-skills so students can practice the ones they missed. Teachers see results by class, student, rubric part and skill.

Scope of the evidence. Magna offers practice for 13 AP courses and TSIA2. The outcome evidence in this report covers two AP courses: AP U.S. Government and Politics (Study 1) and AP Biology (Study 2). Other AP courses and TSIA2 use the same practice-and-feedback loop but have not been studied.

Three lanes of activity feed the same AP exam outcomes

Evidence status of each output and outcome, from Studies 1 and 2

● Significant, $p < .05$ ○ Observed; n.s. or untested ○ Not yet measured



Magna logic model · inputs to long-term outcomes; evidence status from Studies 1 and 2

Theory of change. If students practice free-response writing often, get rubric-specific feedback the moment they submit, work through what they missed with Ask Magna, and build weak sub-skills, while teachers see where the class is stuck, then students close the gap between what they know and what they can write, and more pass and earn 4s and 5s.

Study 1’s FRQ analysis tests the core link directly: the question types practiced in Magna improved more than the one that was not. Skill Builder was introduced after both studies, so its role rests on the logic model and on the planned studies. A full logic model packet, with assumptions, external factors, an annotated bibliography and the research plan, is available from Magna Education.

Research base for the logic model

Each of the six design decisions in the logic model rests on published research on practice, feedback, tutoring, critical thinking and writing. These effects come from other studies and settings: they show why each decision is plausible, while Studies 1 and 2 are the evidence for Magna itself. Effect sizes are standardized mean differences (d or g).

Design decision	What the research shows	Sources
D1 · Frequent practice in AP question formats	Classroom quizzing improved learning across 222 studies and 48,478 students (g = 0.50). An Institute of Education Sciences (IES) guide rates quizzing as strong evidence.	Yang et al., 2021; Pashler et al., 2007
D2 · Feedback scored by rubric part, returned the moment a student submits	In computer-based learning, elaborated feedback outperformed correct/incorrect-only feedback (0.49 vs 0.05). Across 435 studies, feedback averaged d = 0.48. Automated feedback on writing averaged g = 0.55 (secondary students 0.50). With 459 upper-secondary students, GPT-3.5 feedback on essays improved revisions (d = 0.19) and motivation (d = 0.36).	Van der Kleij et al., 2015; Wisniewski et al., 2020; Fleckenstein et al., 2023; Kluger & DeNisi, 1996; Meyer et al., 2024

Design decision	What the research shows	Sources
<p>D3 · An AI tutor students question after feedback (Ask Magna; available in both studies, light use in Study 1)</p>	<p>Students use Ask Magna after they submit and get feedback. It is grounded in that work, stays within course content, and asks guiding questions before giving an answer. Rule-based tutoring systems beat teacher-led large-group instruction ($g = 0.42$); dialogue was linked to larger critical-thinking gains. A research-designed AI tutor beat in-class active learning in a randomized trial (college physics). In high school math, unguarded GPT-4 lowered later exam scores by 17%; a tutor giving hints rather than answers avoided that harm but did not raise exam scores.</p>	<p>Ma et al., 2014; Bastani et al., 2025; Kestin et al., 2025; Abrami et al., 2015</p>
<p>D4 · Questions broken into sub-skills, with practice on the ones missed (Skill Builder, not yet evaluated)</p>	<p>Mastery learning programs averaged 0.52 SD across 108 evaluations, with stronger effects for weaker students. Study 1 found no gain for its lowest-baseline students, so whether Skill Builder helps them is a hypothesis to test. Explicit critical-thinking instruction averaged $g = 0.30$ across 341 effect sizes, with larger gains from dialogue and authentic problems. In grades 6–12, writing strategy instruction averaged 0.76 and feedback 0.34 across 406 comparisons.</p>	<p>Kulik et al., 1990; Graham et al., 2023; Abrami et al., 2015</p>

Design decision	What the research shows	Sources
D5 · Class insights by rubric part, so teachers re-teach	A review of classroom studies found that strengthening frequent formative feedback produces substantial gains. IES rates the evidence for using student data to guide instruction as low, so this is the model's least-supported link.	Black & William, 1998; Hamilton et al., 2009
D6 · Practice built around written explanations of course content	Writing in science, social studies and math improved learning (ES 0.30 across 56 experiments), with similar effects in high school. IES rates asking students deep explanatory questions as strong evidence. Writing instruction in grades 6–12 averaged 0.47.	Graham et al., 2020; Pashler et al., 2007; Graham et al., 2023

Limits of this research base. Feedback can backfire: in one large meta-analysis, over a third of feedback interventions lowered performance (Kluger & DeNisi, 1996), and task-level, elaborated feedback shows the largest effects (Van der Kleij et al., 2015; Wisniewski et al., 2020). Effects tend to be smaller on standardized external tests such as the AP exam (Kulik & Fletcher, 2016; Kulik et al., 1990), and none of the AI studies cited here involve AP courses. Most tutoring evidence comes from rule-based systems rather than chatbots (Ma et al., 2014; Kulik & Fletcher, 2016).

Methods

Both studies compare a classroom's first year with Magna to the same teacher's class the year before, using the College Board AP exam as the outcome.

Settings and participants

Study 1 class. AP U.S. Government and Politics, taught in every year by the same teacher, who has more than ten years of experience with the course: 79 students in the comparison cohort (May 2024 exam) and 62 enrolled, 60 analyzed, in the Magna cohort (May 2025 exam).

Study 1 school. The school is a public, college-preparatory K-12 school in the southern United States with a high school student-teacher ratio of about 10 to 1. Schoolwide in 2024-25, 71.5% of students were White, 16.3% Black, 4.7% Hispanic, 4.4% of two or more races and 2.9% Asian, and 3.4% qualified for free or reduced-price lunch (National Center for Education Statistics). Nearly all graduates enroll in four-year colleges. These are school-level figures; student-level background data were not available for the study cohorts.

Study 2 class. AP Biology, taught by a second teacher, with about 85 students a year: a comparison cohort that took the May 2025 exam and a Magna cohort that took the May 2026 exam.

Instruction before Magna

Study 1. In 2023-24, the comparison year, the teacher taught the course with regular classroom instruction, unit tests, in-class review sessions and free-response practice that the teacher wrote and graded by hand. Grading time limited how much free-response practice could be assigned. In 2024-25 all of this continued unchanged, and Magna added AI-scored practice on top of it.

In Study 2, the 2024-25 AP Biology class was taught without Magna. In 2025-26 the teacher added Magna; because Magna questions were not always ready ahead of each unit, use was lower than planned. Ask Magna use began in December 2025: 48 of the 84 students on the Magna class roster (57%) used it at least once, with 30 using it in December, 6 to 13 in each month from January to May, and 1 in June. Use peaked in December and again in April, before the exam.

Study selection

These two classrooms are the only Magna classrooms that met both inclusion criteria: a pass rate (3 or higher) below 85% in the year before Magna, and a comparable unit test given before Magna use in both the comparison year and the Magna year. Both criteria

depend only on pre-intervention data, not on outcomes. Classrooms already passing at 85% or more leave little room for measurable change on the AP exam, and a comparable pre-intervention test is needed to account for differences between cohorts.

Design and analysis

Study 1 uses student-level data and adjusts each student's AP outcome for the Unit 1 test taken before Magna use. Study 2 uses class-level results and compares the class's change with the change among all AP Biology test takers. Model details appear with each study's results below.

Study 1: AP U.S. Government and Politics

After adjustment for pre-intervention achievement, the Magna cohort had higher odds of scoring 4 or 5 (the primary outcome), higher odds of scoring 5 and a higher mean AP score than the same teacher's prior-year cohort.

Design. A quasi-experimental historical-cohort study using a nonequivalent comparison group and student-level statistical adjustment for pre-intervention achievement. Keeping the teacher constant holds teacher-level differences fixed, a major source of variation in historical comparisons.

Intervention. Magna use began in spring 2025, after the Unit 1 test, and continued through the May 2025 exam; curriculum, pacing and class time were unchanged. Students practiced AP-style multiple-choice items and three of the four free-response (FRQ) types: Concept Application, Quantitative Analysis and SCOTUS Comparison. Magna scored each FRQ attempt against the College Board rubric and returned immediate feedback. The fourth type, the Argument Essay, was not supported in Magna.

Implementation. The 2024–25 Magna class report includes all 62 enrolled students, with 277 scored attempts on four class-wide assessments (two multiple-choice, two FRQ). Between 52 and 60 students completed each assessment, with a median of 9.6 minutes per assessment (interquartile range 7.5 to 12.4). Practice beyond these assessments was not logged per student. Ask Magna use was logged by class and month: 19 of the 62 students (31%) used it at least once during the year, and 14, 3 and 6 students used it in February, March and April 2025, so Ask Magna use in Study 1 was light.

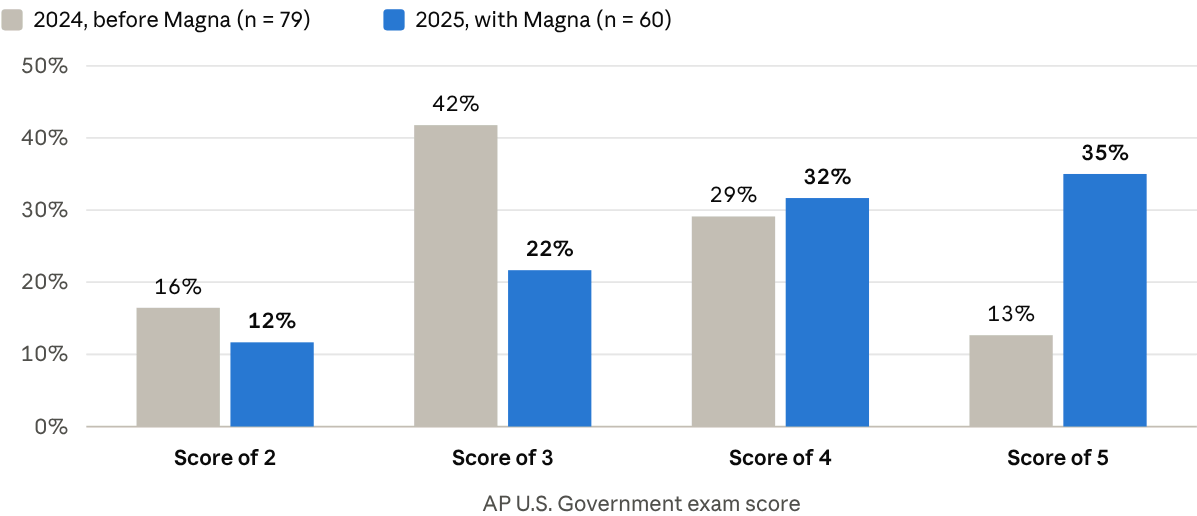
Measures. The pre-intervention measure is the teacher's Unit 1 test (30 points): mean 20.72 (SD 3.17) in 2024 and 21.90 (SD 2.81) in 2025, a difference of 1.18 points ($p = .022$, $d = 0.39$). Unit 1 scores correlated with AP scores at $r = 0.46$ (2024) and $r = 0.66$ (2025). The

outcome is the College Board AP score (1 to 5).

Analysis. Because the 2025 cohort entered with somewhat higher pre-intervention achievement, all primary student-level models adjusted for Unit 1 score to reduce selection bias associated with differences in incoming academic performance. All models pool the 139 students, with Magna cohort (0/1) and Unit 1 score as predictors. Scoring 4 or higher (logistic regression) is the primary outcome; it was the primary outcome of the teacher’s original 2025 analysis, which modeled scores of 4 or higher and of 5, and this reanalysis kept it. Secondary outcomes are scoring 5 and 3 or higher (logistic regression), the full score distribution (ordinal logistic regression) and mean AP score (linear regression with HC3 robust standard errors). The FRQ comparison, moderation analysis and follow-up year are exploratory.

The share of 5s nearly tripled as students moved up from 3s

Share of students at each AP score (%). No student scored 1 in either year.



Student-level College Board AP scores, 2024 (n = 79) and 2025 (n = 60)

Outcome, 2025 cohort	Predicted without Magna*	Observed	Adjusted effect (95% CI)	p
Scored 4 or 5 (primary)	50%	67%	Odds ratio 2.25 (1.04 to 4.87)	.039
Scored 5	17%	35%	Odds ratio 3.25 (1.25 to 8.49)	.016
Mean AP score	3.58	3.90	+0.32 points (0.03 to 0.61)	.032

Outcome, 2025 cohort	Predicted without Magna*	Observed	Adjusted effect (95% CI)	p
Full score distribution	—	—	Ordinal odds ratio 2.00 (1.04 to 3.86)	—
Scored 3 or higher	89%	88%	Odds ratio 0.91 (0.29 to 2.83)	.86

*Each model's prediction for the 2025 students' own Unit 1 scores with the Magna term set to zero.

The Magna cohort did not significantly increase the already-high 3+ rate (83.5% in the comparison cohort); the observed effect was concentrated in students moving upward into scores of 4 and 5.

Standardized effect sizes. Following What Works Clearinghouse conventions, the adjusted effect on mean AP score is Hedges' $g = 0.33$ (0.03 to 0.63): the adjusted difference divided by the pooled standard deviation of AP scores (0.96). For the binary outcomes, the Cox index (log odds ratio divided by 1.65, with the small-sample correction) gives $g = 0.49$ (0.02 to 0.95) for scoring 4 or higher, 0.71 (0.13 to 1.29) for scoring 5 and -0.06 (-0.75 to 0.63) for scoring 3 or higher. Unadjusted, the difference in mean AP score is $g = 0.54$; the difference in Unit 1 scores before Magna was $g = 0.39$.

Robustness. The unadjusted mean rose from 3.38 to 3.90 (Welch $p = .002$). Students matched exactly on Unit 1 score (58 of 60) differ by +0.49 points. A 5,000-draw permutation test of the adjusted effect gives $p = .028$. As context rather than as a control, national results over the same years show the share scoring 4 or 5 moving from 49.3% to 48.5% and the mean from 3.38 to 3.34, so the gain was not part of a nationwide rise.

Nonequivalent dependent variable (exploratory). The FRQ types practiced in Magna were compared with the Argument Essay, which was not, using each question's gap to the national benchmark in percent of maximum points (difference-in-differences, 2024 to 2025). The practiced types gained more than the essay by an average of +10.7 points against the global benchmark and +11.2 against the state benchmark.

FRQ type practiced in Magna	Gain beyond the essay, vs global (% of max)	95% interval†
Concept Application	+11.5	-1.7 to +24.8
Quantitative Analysis	+7.1	-4.6 to +18.8

FRQ type practiced in Magna	Gain beyond the essay, vs global (% of max)	95% interval [†]
SCOTUS Comparison	+13.7	+0.7 to +26.8
Average of three	+10.7	+0.5 to +21.1

[†]College Board class reports give class means by rubric part, not student-level FRQ scores. Intervals are simulated from part-level means with sampling and rounding error, treating parts as independent within students; if parts were perfectly correlated, the average's interval would be -6.6 to +28.1.

Moderation by baseline (exploratory). The effect increased with pre-intervention achievement (Unit 1 × Magna interaction, $p = .013$). Estimated gains were +0.31 points at Unit 1 scores of 19 to 21, +0.56 at 22 to 23 and +0.67 at 24 to 28. No gain was observed in the lowest band (18 or below; 14 and 7 students).

Follow-up year (descriptive). In Magna's second year, the 2026 cohort ($n = 78$) had 59% scoring 4 or higher and 87% scoring 3 or higher, against 51% and 76% of all test takers and 41.8% (4+) in the 2024 comparison cohort. Unit 1 scores were not collected for 2026, so this cohort is not adjusted. Only 7 students in this class used Ask Magna during the year.

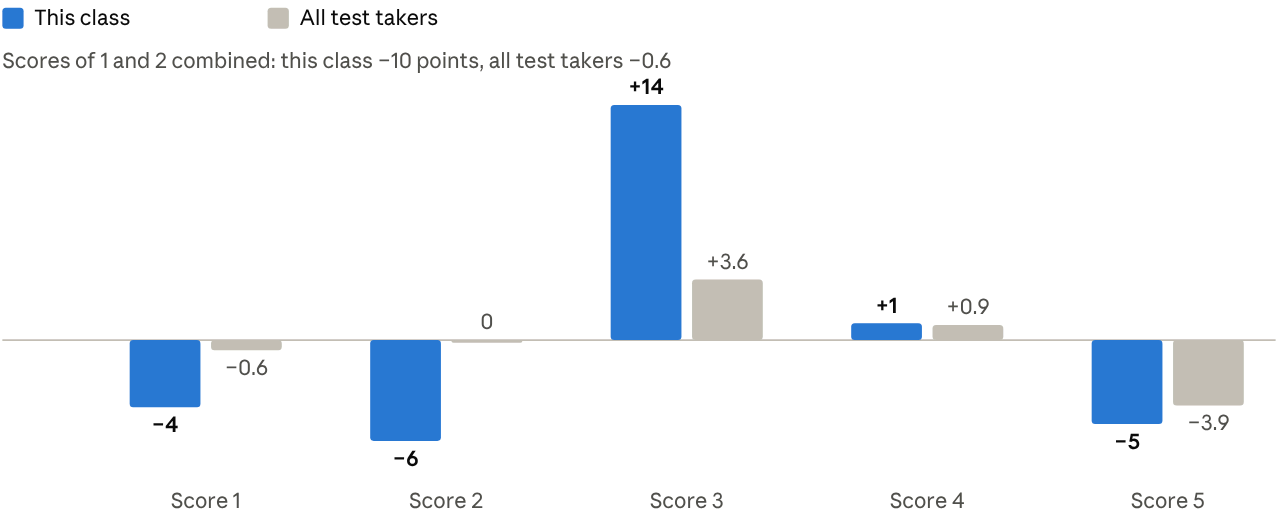
Study 2: AP Biology (supporting replication)

Study 2 is presented as supporting replication across a second teacher and subject, not as the primary qualifying study. With equivalent pre-intervention scores, the AP Biology pass rate rose 10.1 percentage points in Magna's first year while the national rate rose 0.8, and the share of students scoring 1 or 2 fell by about half. These results do not reach conventional statistical significance on their own. On Study 1's primary outcome, scoring 4 or higher, AP Biology did not improve: the share fell from 54% to 50% (about 1 point net of the national change).

Setting and measures. AP Biology taught by a second teacher, with about 85 students a year. The comparison cohort took the May 2025 exam and the Magna cohort the May 2026 exam. The pre-intervention measure is the teacher's first unit test, given before Magna use: class means of 76.9% (2025) and 76.7% (2026). Outcomes are College Board AP scores, compared with the change among all AP Biology test takers. Pass rate is the main outcome because it is the measure the teacher reported; the planned study pre-registers the pass rate and the share scoring 4 or 5 as co-primary outcomes.

The share of 1s and 2s fell 10 points in class while national shares held steady

Change in share of AP Biology students at each score, 2025 to 2026 (percentage points)



Class: teacher-reported score distribution, 2025 and 2026 (about 88 and 87 students) · All test takers: College Board AP Biology score distributions, 2026 rounded

Outcome	2025 cohort	2026 Magna cohort	National change	Change net of national (95% CI)	p
Pass rate (3 or higher)	78.4%	88.5%	70.4% to 71.2%	+9.3 points; odds ratio 2.04 (0.89 to 4.69)	.09
Scored 1 or 2	21.6%	11.5%	29.6% to 29%	-9.5 points	.10*
Scored 4 or 5	54%	50%	43.0% to 40%	About -1 point	.65*
Mean AP score	3.49	3.54	3.24 to 3.19	+0.10 (-0.21 to +0.41)	.53

*Fisher exact test of the class change. The share scoring 5 fell 5 points in class and 3.9 points nationally.

Educator and student voices

Study 1 teacher. The teacher reported that automatic scoring saved time in both creating and grading free-response assessments, which made it possible to assign students more free-response practice than before.

Study 2 teacher and students. The teacher reports that Magna questions were not always ready ahead of each unit in 2025-26, so use was lower than planned and a few students did not use it. In the teacher's end-of-year conversations, many students

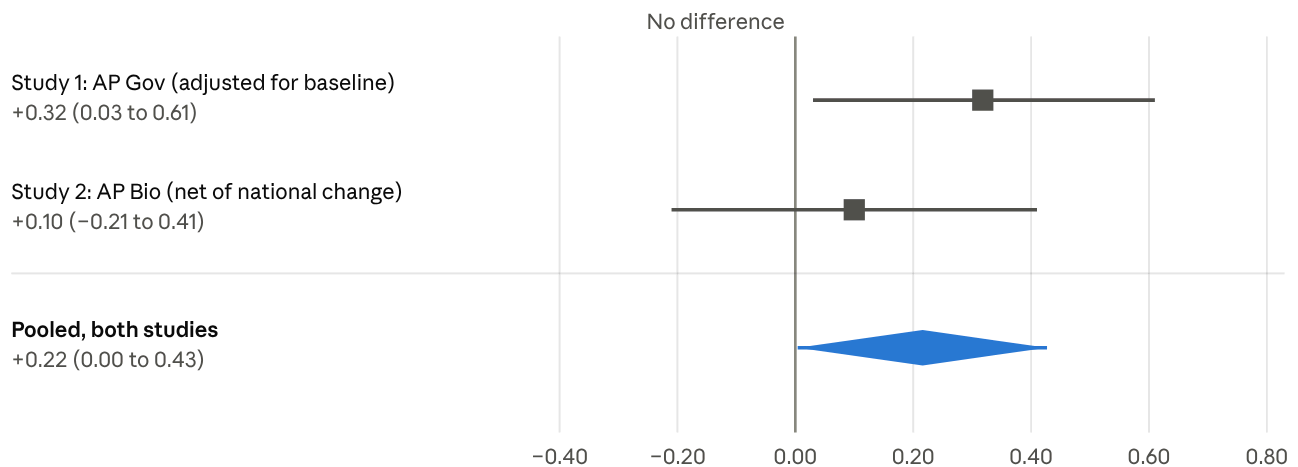
described Magna positively and said its explanations were clearer than those of general-purpose chatbots. Students cited the practice tests, the Ask Magna tutor, and prompts asking them to reflect on their own understanding.

Exploratory cross-study synthesis

This synthesis is descriptive and is not part of the Tier 3 case, because the two studies differ in data quality and analytic structure. Combining them, the Magna cohorts scored an estimated +0.22 points above expectation on mean AP score (95% CI 0.00 to 0.43), with no significant difference between studies (heterogeneity $p = .31$).

Exploratory synthesis: Magna cohorts scored about 0.22 points above expectation

Effect on mean AP score in Magna's first year, points, with 95% confidence interval



Study 1: baseline-adjusted linear model, $n = 139$ · Study 2: class change minus national change, $n \approx 88$ and 87 · Pooled: fixed-effect inverse-variance average

The pooled estimate is a fixed-effect inverse-variance average. Study 1's estimate is adjusted student by student for baseline. Study 2's is the class change net of the national change, with equivalent class baselines.

ESSA Tier 3 alignment

Study 1 is the qualifying study, with scoring 4 or higher as its primary outcome. Study 2 is presented as supporting replication across a second teacher and subject.

Tier 3 criterion (U.S. Department of Education, 2023)	Evidence from Study 1
Correlational study with statistical controls for selection bias	Quasi-experimental historical-cohort design with a nonequivalent comparison group: the same teacher's prior-year class. Because the 2025 cohort entered with somewhat higher pre-intervention achievement, all primary student-level models adjusted for Unit 1 score to reduce selection bias associated with differences in incoming academic performance. Exact baseline matching and a 5,000-draw permutation test gave consistent results.
Statistically significant favorable effect	Primary outcome, scoring 4 or higher: odds ratio 2.25 (1.04 to 4.87), $p = .039$. Secondary outcomes: scoring 5, $p = .016$; mean AP score, $p = .032$.
Relevant student outcome	College Board AP exam scores, scored independently of the developer and the teacher. Scores of 4 and 5 represent stronger AP performance than a 3, and at many colleges they qualify for credit or placement that a 3 does not; policies vary by institution.
No overriding statistically significant negative findings	No statistically significant adverse outcome was observed in either study. In Study 1 the 3+ result was nonsignificant (odds ratio 0.91, $p = .86$), while the 4+, 5 and mean-score outcomes were significantly favorable. In Study 2 the share scoring 4 or 5 changed by about 1 point net of the national change ($p = .65$).

Tier 3 criterion (U.S. Department of Education, 2023)	Evidence from Study 1
Well designed and implemented	The same teacher taught both cohorts, which holds teacher-level differences constant. Curriculum, pacing and class time were unchanged, and Magna was added to instruction rather than replacing it. Platform records include all 62 enrolled students. Attrition was 2 of 62 (3%), both relocations.
Consistency with other findings	Study 2, with a second teacher, subject and exam year, showed a non-significant rise in the pass rate ($p \approx .09$), and its share scoring 4 or 5 fell from 54% to 50% (about 1 point net of the national change, not significant). The 2026 follow-up cohort in Study 1 continued to score above the national distribution.

Study 1 shows statistically significant favorable effects on relevant student outcomes and no statistically significant adverse effects. It therefore provides evidence consistent with the requirements for ESSA Tier 3 (Promising Evidence).

Components covered. This evidence applies to the Magna components used during the study period: AP-style practice, automated rubric scoring, immediate feedback and the Ask Magna AI tutor, which was available to students in both studies. Ask Magna use was recorded by class and month, not linked to individual outcomes, so its separate contribution cannot be estimated; 19 of 62 students used it in Study 1 and 48 of 84 in Study 2. Skill Builder was introduced after both studies and was not evaluated.

Conclusion

In one AP U.S. Government classroom, students in the year Magna was added were more likely to score 4 or 5 than the same teacher’s previous class, after accounting for how each student started (odds ratio 2.25, $g = 0.49$). The free-response types students practiced in Magna improved more than the type they did not, consistent with the program’s theory of change. A second classroom, in AP Biology, saw a non-significant rise in its pass rate and no rise in scores of 4 and 5. Together, the studies are promising evidence for Magna’s practice-and-feedback loop in AP courses.

Implementation lessons.

- Add Magna on top of existing instruction rather than in place of it, as both classrooms did.
- Have Magna questions ready before each unit is taught; the Study 2 teacher reported lower use when they were not.
- Give a pre-test before Magna use and track student use, so a school can measure its own results.

The planned studies, with same-year comparison classrooms, pre-registered co-primary outcomes (pass rate and scores of 4 or 5) and an independent evaluator, are designed to test these findings more rigorously.

Limitations and next studies

These findings should be read as promising rather than confirmatory evidence.

- **Historical comparison.** The comparison group is the prior-year cohort rather than a concurrent group.
- **Year and intervention coincide.** Because Magna exposure coincided with school year, the design cannot fully rule out unmeasured year-to-year differences between cohorts. The same teacher, unchanged instructional conditions, baseline adjustment, national benchmarks, matching and permutation analysis reduce, but do not eliminate, this concern.
- **One teacher and one school** for the qualifying study.
- **Selection on a prior-year threshold.** Classrooms were included because their pass rate before Magna was below 85%, which can by itself make the following year look better (regression to the mean). Earlier years of results for these classrooms were not available to test this.
- **Exam change.** The AP U.S. Government exam moved to a fully digital format in the Magna year. National benchmarks reflect the change but cannot rule out a classroom-specific effect.
- **Limited covariates.** Unit 1 score is the only pre-intervention measure; no student-level demographic data were available.
- **Retrospective, developer-conducted analysis.** The Study 1 analysis was retrospective, its outcomes were not pre-registered, and Magna Education conducted it.

- **Usage data.** Platform records cover class-wide assessments and monthly class totals for Ask Magna; other practice was not logged per student, so dosage cannot yet be linked to outcomes.
- **Data granularity.** Study 2 uses class-level results, with class sizes estimated from the reported pass rates (conclusions are unchanged for any size from 70 to 100 per year). The FRQ analysis in Study 1 uses class means by rubric part.
- **Setting.** The qualifying study took place in a college-preparatory school with few economically disadvantaged students. Results may differ in schools with more disadvantaged students or lower AP pass rates.

Planned next studies (2026–27 or 2027–28)

The evaluation will extend to additional classrooms and courses with a baseline test before Magna use, per-student usage logs from the Magna platform, student-level AP score rosters, and two pre-registered co-primary outcomes, the pass rate (share scoring 3 or higher) and the share scoring 4 or 5, since the theory of change predicts both, with a correction for testing two outcomes. The plan uses same-year comparison classrooms, at least two teachers and 30 students per condition, and analysis by an independent evaluator rather than Magna. Usage logs will allow a dose-response analysis. These prospective studies are designed to strengthen causal inference and to evaluate treatment dosage and implementation fidelity.

References

Study data and benchmarks

- College Board AP U.S. Government and Politics student scores and class reports, 2024–2026
- College Board AP Biology class results, 2025 and 2026
- Magna Education platform records: Ask Magna use by class and month, 2024–25 and 2025–26
- [College Board: AP Biology score distributions](#)
- College Board published AP U.S. Government and Politics score distributions, 2024–2026
- [College Board: AP Biology updates for 2025–26](#)
- National Center for Education Statistics, Common Core of Data, 2024–25 (school-level demographics)

- [U.S. Department of Education: Selecting evidence-based practices for Tiers 1, 2 and 3](#)
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

National AP free-response data

- College Board. (2025). AP U.S. Government and Politics 2025 Chief Reader Report, [Set 1](#) and [Set 2](#).
- College Board. (2024). [AP U.S. Government and Politics 2024 Chief Reader Report, Set 1](#).
- College Board. (2025). [AP Biology 2025 Chief Reader Report](#).

Research on practice, feedback, tutoring and writing

- Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. *Review of Educational Research*, 85(2), 275–314. <https://doi.org/10.3102/0034654314551063>
- 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), e2422633122. <https://doi.org/10.1073/pnas.2422633122>
- Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74. <https://doi.org/10.1080/0969595980050102>
- Fleckenstein, J., Liebenow, L. W., & Meyer, J. (2023). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence*, 6, 1162454. <https://doi.org/10.3389/frai.2023.1162454>
- Graham, S., Kim, Y.-S., Cao, Y., Lee, J. W., Tate, T., Collins, P., Cho, M., Moon, Y., Chung, H. Q., & Olson, C. B. (2023). A meta-analysis of writing treatments for students in grades 6–12. *Journal of Educational Psychology*, 115(7), 1004–1027. <https://doi.org/10.1037/edu0000819>
- Graham, S., Kiuvara, S. A., & MacKay, M. (2020). The effects of writing on learning in science, social studies, and mathematics: A meta-analysis. *Review of Educational Research*, 90(2), 179–226. <https://doi.org/10.3102/0034654320914744>
- Hamilton, L., Halverson, R., Jackson, S., Mandinach, E., Supovitz, J., & Wayman, J. (2009). *Using student achievement data to support instructional decision making* (NCEE 2009-4067). Institute of Education Sciences, U.S. Department of Education.

- Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15, 17458. <https://doi.org/10.1038/s41598-025-97652-6>
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284. <https://doi.org/10.1037/0033-2909.119.2.254>
- Kulik, C.-L. C., Kulik, J. A., & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research*, 60(2), 265–299. <https://doi.org/10.3102/00346543060002265>
- Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, 86(1), 42–78. <https://doi.org/10.3102/0034654315581420>
- Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, 106(4), 901–918. <https://doi.org/10.1037/a0037123>
- Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M. H., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 6, 100199. <https://doi.org/10.1016/j.caeai.2023.100199>
- Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning* (NCER 2007-2004). Institute of Education Sciences, U.S. Department of Education.
- Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research*, 85(4), 475–511. <https://doi.org/10.3102/0034654314564881>
- Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, 10, 3087. <https://doi.org/10.3389/fpsyg.2019.03087>
- Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*, 147(4), 399–435. <https://doi.org/10.1037/bul0000309>

About this report

Authorship and analysis. This report was prepared by Magna Education. The Study 1 teacher designed the original comparison and collected the data; Magna Education re-analyzed the student-level data and College Board reports in October 2026. The Study 2 teacher supplied class-level results. The Study 1 teacher and school were not compensated for the study .

Recommended citation. Magna Education. (2026). *Magna Education and AP exam outcomes: Study report*. San Francisco, CA: Magna Education.

Contact. hello@magnaeducation.ai · magnaeducation.ai