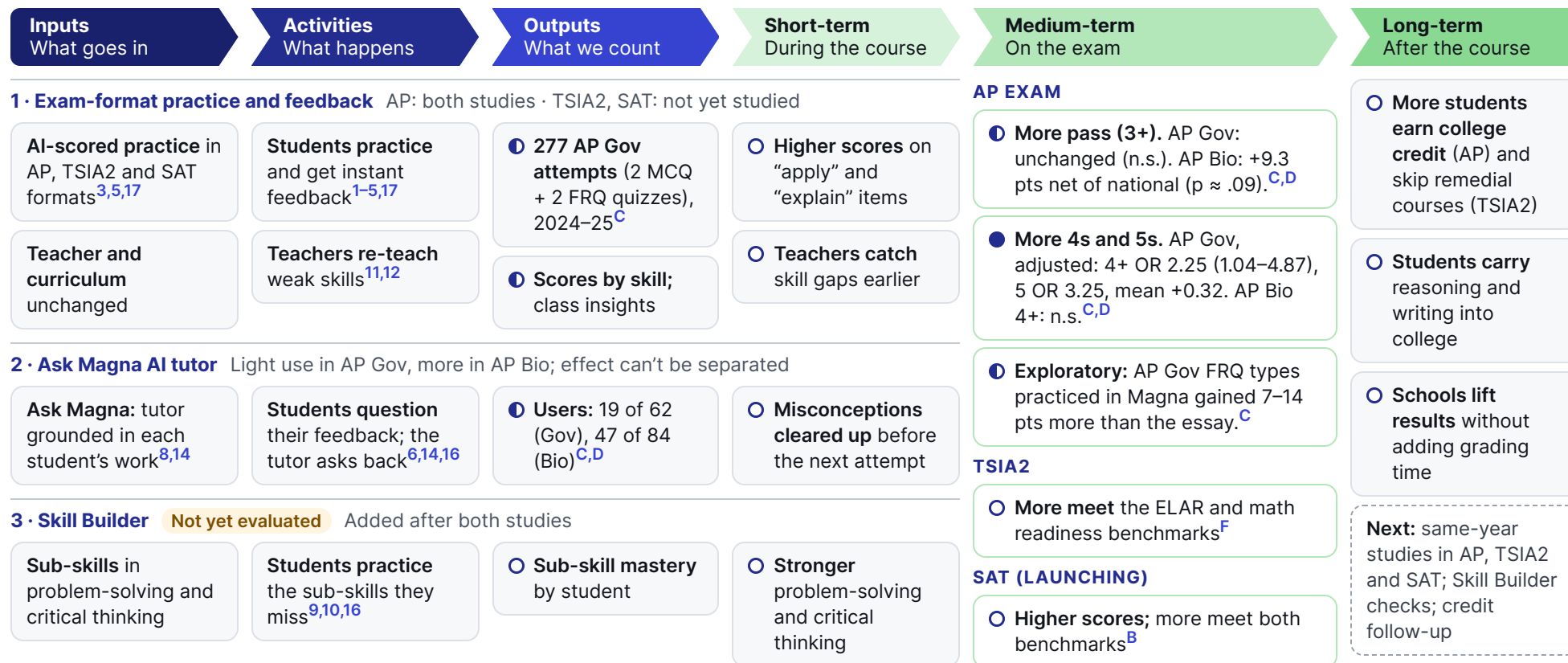


# Magna logic model: AP, TSIA2 and SAT

**Theory of change:** if students practice in each exam's format often, get feedback the moment they submit, work through mistakes with Ask Magna and rebuild weak sub-skills, while teachers see where the class is stuck, then they master the tested skills and more reach each exam's benchmarks.

## THE PROBLEM WE TARGET (NATIONAL DATA)

On a 2025 AP Gov free-response question, students earned 94% of "identify" but 39% of "explain" points.<sup>A</sup> Only 39% of 2025 SAT takers met both readiness benchmarks.<sup>B</sup> 43% of 2023 Texas public college entrants were not college ready in at least one subject.<sup>E</sup>



### ASSUMPTIONS

- Students complete practice and use Ask Magna
- Explained feedback builds tested skills<sup>3-5,15</sup>
- Practice and AI scores match each exam
- Teachers act on the class insights

### EXTERNAL FACTORS

- Differences between cohorts and years
- Teacher experience and course pacing
- Exam, scoring or benchmark changes
- Outside test prep; test-optional policies

### RESEARCH PLAN

- Same-year comparison, 2+ teachers per arm
- Common pretest; gap under 0.25 SD
- Pre-registered outcomes for each exam
- Independent analysis; per-student usage logs

**Evidence:** AP US Gov (Study 1): one teacher, 2024 (n=79) vs 2025 (n=60); retrospective, analyzed by Magna; class chosen for a pre-Magna pass rate under 85%; Magna cohort started 0.39 SD higher; exam went digital in 2025; associations, not causal estimates. AP Bio (Study 2): second teacher, class-level, 2025 vs 2026, equal Unit 1 baselines. TSIA2 and SAT: not yet studied. **Scope:** 13 AP courses and TSIA2 offered; SAT practice launching. Platform version 1.0 · October 2026 · page 1 of 4

## Six design decisions and the research behind each

Each design decision maps to a lane on page 1. The effects below come from other studies: they show why each decision is plausible, not Magna's own effect (d or g = standardized mean difference). Annotated sources: pages 3–4.

| DESIGN DECISION                                                                                                              | WHAT THE RESEARCH SHOWS                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | SOURCES                                                                                                                                       |
|------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------|
| <b>D1 · Frequent practice in each exam's question formats</b><br>Lane 1 · activity                                           | 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.                                                                                                                                                                                                                                                                                                                                                                                                             | 1 Yang et al., 2021<br>2 Pashler et al., 2007                                                                                                 |
| <b>D2 · Instant feedback, by rubric part or with an explained answer</b><br>Lane 1 · input and activity                      | 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).                                                                                                                                                                                                                        | 3 Van der Kleij et al., 2015<br>4 Wisniewski et al., 2020<br>5 Fleckenstein et al., 2023<br>15 Kluger & DeNisi, 1996<br>17 Meyer et al., 2024 |
| <b>D3 · An AI tutor students question after feedback (Ask Magna)</b><br>Lane 2 · in both studies; light use in AP Gov        | 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. | 6 Ma et al., 2014<br>8 Bastani et al., 2025<br>14 Kestin et al., 2025<br>16 Abrami et al., 2015                                               |
| <b>D4 · Questions broken into sub-skills, with practice on the ones missed (Skill Builder)</b><br>Lane 3 · not yet evaluated | 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. Critical-thinking instruction averaged g = 0.30 across 341 effect sizes; writing strategy instruction averaged 0.76 in grades 6–12.                                                                                                                                                                                                                | 9 Kulik et al., 1990<br>10 Graham et al., 2023<br>16 Abrami et al., 2015                                                                      |
| <b>D5 · Class insights by skill and rubric part, so teachers re-teach</b><br>Lane 1 · activity                               | 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.                                                                                                                                                                                                                                                                                                                                                | 11 Black & Wiliam, 1998<br>12 Hamilton et al., 2009                                                                                           |
| <b>D6 · Practice built around explaining reasoning</b><br>Lane 1 · FRQs, TSIA2 essay                                         | 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.                                                                                                                                                                                                                                                                                                                        | 13 Graham et al., 2020<br>2 Pashler et al., 2007<br>10 Graham et al., 2023                                                                    |

### WHAT THE RESEARCH ALSO WARNS

- Over a third of feedback interventions lowered performance (15).
- Tutoring evidence is mostly rule-based; unguarded chatbots can hurt (6–8).
- Effects shrink on standardized tests like AP, TSIA2, SAT (7, 9).
- Commercial SAT coaching has raised scores only modestly (18).
- None of the AI studies cited here involve AP, TSIA2 or the SAT.

### WHAT MAGNA'S OWN EVIDENCE SHOWS

- AP Gov, adjusted: 4+, 5s and mean rose (n = 139).
- AP Gov: no gain in the 3+ rate or the lowest-baseline band.
- AP Bio: no significant change; the 4+ share fell (54% → 50%).
- Ask Magna: used by 19 of 62 (Gov), 47 of 84 (Bio); not separable.
- Not yet studied: TSIA2, SAT, other AP courses, Skill Builder.

# Annotated bibliography

Each source: what it studied, what it found, and the design decisions (D1–D6, page 2) it informs.

- 1 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. [doi.org/10.1037/bul0000309](https://doi.org/10.1037/bul0000309)  
Meta-analysis of 222 classroom studies and 48,478 students. Quizzing improved learning compared with other conditions ( $g = 0.50$ ). **Informs D1.**
- 2 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.  
IES practice guide. Rates using quizzes to re-expose students to key content (Recommendation 5b) and asking deep explanatory questions (Recommendation 7) as strong evidence. **Informs D1, D6.**
- 3 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. [doi.org/10.3102/0034654314564881](https://doi.org/10.3102/0034654314564881)  
Meta-analysis of 40 studies (70 effect sizes) in computer-based settings. Elaborated feedback (0.49) beat giving the correct answer (0.32) and right/wrong only (0.05), especially for higher-order outcomes. **Informs D2.**
- 4 Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, 10, 3087. [doi.org/10.3389/fpsyg.2019.03087](https://doi.org/10.3389/fpsyg.2019.03087)  
Meta-analysis of 435 studies (more than 61,000 students). Feedback averaged  $d = 0.48$ ; high-information feedback  $d = 0.99$  (42 effect sizes); reinforcement or punishment  $d = 0.24$ . **Informs D2.**
- 5 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. [doi.org/10.3389/frai.2023.1162454](https://doi.org/10.3389/frai.2023.1162454)  
Meta-analysis of 20 studies (84 effect sizes, 2,828 students). Automated feedback on writing averaged  $g = 0.55$ , and 0.50 for secondary students, with significant variation across studies. **Informs D2.**
- 6 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. [doi.org/10.1037/a0037123](https://doi.org/10.1037/a0037123)  
Intelligent tutoring systems outperformed teacher-led large-group instruction ( $g = 0.42$ ) and other computer instruction (0.57), and did not differ significantly from one-to-one human tutoring. These were rule-based systems, not chatbots. **Informs D3 and the page 2 cautions.**
- 7 Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, 86(1), 42–78. [doi.org/10.3102/0034654315581420](https://doi.org/10.3102/0034654315581420)  
Across 50 evaluations, tutoring systems raised test scores by a median 0.66 SD. Effects were larger on locally developed tests than on standardized ones, a caution for AP outcomes. **Informs the page 2 cautions.**
- 8 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. [doi.org/10.1073/pnas.2422633122](https://doi.org/10.1073/pnas.2422633122)  
Field experiment with about 1,000 high school students. Plain GPT-4 raised practice grades but lowered later unassisted exam scores by 17%. A tutor that gave hints rather than answers avoided the harm but did not raise exam scores. **Informs D3 and the page 2 cautions.**
- 9 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. [doi.org/10.3102/00346543060002265](https://doi.org/10.3102/00346543060002265)  
Meta-analysis of 108 evaluations. Mastery learning raised exam scores by 0.52 SD, with stronger effects for weaker students and smaller effects on standardized tests. These were classroom mastery-learning programs, not software. **Informs D4 and the page 2 cautions.**
- 10 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. [doi.org/10.1037/edu0000819](https://doi.org/10.1037/edu0000819)  
406 comparisons, 52,529 students in grades 6–12. Writing instruction averaged 0.47; strategy instruction 0.76; feedback 0.34. **Informs D4, D6.**

## Annotated bibliography, continued

- 11 Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74.  
[doi.org/10.1080/0969595980050102](https://doi.org/10.1080/0969595980050102)  
Narrative review of classroom research: strengthening the frequent feedback students receive yields substantial learning gains. **Inform**s D5.
- 12 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.  
IES practice guide on teachers' use of student data. All five recommendations are rated low evidence, so D5 is the model's least-supported link. **Inform**s D5.
- 13 Graham, S., Kiuhara, 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. [doi.org/10.3102/0034654320914744](https://doi.org/10.3102/0034654320914744)  
56 experiments. Writing about content improved learning (ES 0.30), with similar effects in elementary, middle and high school. **Inform**s D6.
- 14 Kestin, G., Miller, K., Kiales, 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. [doi.org/10.1038/s41598-025-97652-6](https://doi.org/10.1038/s41598-025-97652-6)  
Randomized trial in a college physics course (194 students). An AI tutor built on the same research-based teaching practices as the class helped students learn significantly more in less time than in-class active learning. Evidence for a designed tutor, not a general chatbot. **Inform**s D3.
- 15 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. [doi.org/10.1037/0033-2909.119.2.254](https://doi.org/10.1037/0033-2909.119.2.254)  
607 effect sizes. Feedback helped on average ( $d = 0.41$ ), but over a third of interventions lowered performance, which is why D2 uses task-level, elaborated feedback. **Inform**s D2 and the page 2 cautions.
- 16 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. [doi.org/10.3102/0034654314551063](https://doi.org/10.3102/0034654314551063)  
341 effect sizes from quasi- and true-experimental studies. Critical-thinking instruction averaged  $g = 0.30$ ; dialogue, authentic problems and mentoring were linked to larger gains. **Inform**s D3, D4.
- 17 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. [doi.org/10.1016/j.caeai.2023.100199](https://doi.org/10.1016/j.caeai.2023.100199)  
459 upper-secondary students wrote argumentative essays in English as a foreign language. Compared with no feedback, GPT-3.5 feedback improved revisions ( $d = 0.19$ ) and task motivation ( $d = 0.36$ ). **Inform**s D2.
- 18 Briggs, D. C. (2009). *Preparation for college admission exams* (NACAC Discussion Paper). National Association for College Admission Counseling.  
Review of research on commercial coaching for the SAT and ACT. Coaching raised SAT scores by about 10–20 points in math and 5–10 points in critical reading on average (200–800 section scale). **Inform**s the page 2 cautions.

### DATA SOURCES (LETTERED)

- A College Board. (2025). *AP U.S. Government and Politics 2025 Chief Reader Report*, Set 1 and Set 2. Question 2 part means (each part worth 1 point): Set 1 A 0.96, B 0.95, C 0.38, D 0.36; Set 2 A 0.94, B 0.89, C 0.49, D 0.32. Identify/describe =  $(0.96 + 0.95 + 0.94 + 0.89) / 4 = 93.5\%$ . Explain/conclude =  $(0.38 + 0.36 + 0.49 + 0.32) / 4 = 38.8\%$ . Unweighted across the two sets.
- B College Board. (2025). *SAT Suite of Assessments annual report: Total group*. Class of 2025: 2,004,965 SAT takers; 39% met both college-readiness benchmarks (ERW 480, Math 530), 64% ERW, 41% Math. [reports.collegeboard.org](https://reports.collegeboard.org)
- C Magna Education. AP U.S. Government and Politics study data: College Board AP scores, 2024 and 2025 cohorts; 2024–25 Class Insights Report; Ask Magna use by month.
- D Magna Education. AP Biology class results, 2025 and 2026; Ask Magna use, 2025–26.
- E Texas Higher Education Coordinating Board. (2025, January). *Supporting underprepared students: Continued progress and recommendations*. Fall 2023: 89,467 of 208,741 first-time-in-college students (43%) were not college ready in at least one subject. [reportcenter.highered.texas.gov](https://reportcenter.highered.texas.gov)
- F Texas Education Agency. *TSIA (Texas Success Initiative Assessment)*. College-readiness benchmarks: Math 950, or below 950 with diagnostic level 6; ELAR 945 with an essay score of 5–8, or below 945 with diagnostic level 5–6 and an essay of 5–8. [tea.texas.gov](https://tea.texas.gov)