The Sethi Ledger

PEOPLE / PRACTICE / POSSIBILITY

LEADERSHIP • WEEKEND RESEARCH SUPPLEMENT

# What students can do without AI

September 26, 2026 to October 2, 2026

This week's research asks how to distinguish confidence, assisted performance and independent understanding. Classroom studies, assessment analyses and interface research offer practical tests for district pilots, with clear limits on what their findings establish.

## The week at a glance

**01** Measure confidence and demonstrated reasoning separately. A well-designed AI activity can still produce validation-seeking rather than deeper dialogue.

**02** Review the whole instructional design. A classifier, a rubric or a verified answer does not explain a learning outcome on its own.

**03** Protect access and agency. Use uncertain predictions to invite participation, and ensure monitoring does not erase work or create unnecessary barriers.

**04** Test transfer after support is removed. Later, unaided performance is a stronger adoption signal than successful completion inside a tool.

### ABOUT THIS BRIEF

Selected research published September 26, 2026 to October 2, 2026. Findings are assessed for relevance to K-12 leadership, evidence quality and practical implications. Each summary identifies important limitations.

01 / LEADERSHIP

# Confidence is not critical thinking

Preprint / mixed-methods pilot
Zahra and colleagues / arXiv

### THE FINDING

Can a chatbot become a thinking partner rather than an answer supplier? A semester-long pilot followed 14 undergraduates in an applied statistics course that asked them to think independently, consult ChatGPT, then explain how their thinking changed. Students reported greater statistical confidence and stronger AI literacy in eight of nine dimensions. Yet the standardized critical-thinking measures showed no statistically significant improvement. Analysis of their interactions also found that students commonly sought validation rather than sustained dialogue. This distinction matters: the positive statistics results measured perceived mastery, not an independent achievement test. The pilot suggests that a thoughtful instructional protocol can coexist with shallow AI use; it does not establish either a learning benefit or an absence of benefit.

### THE EVIDENCE

The preprint combines before-and-after questionnaires, Watson-Glaser reasoning measures, observations and 442 student prompts. There was no comparison class. The statistics confidence instrument developed for this study showed post-test ceiling effects.

### LIMITS TO KEEP IN VIEW

Small, self-selected university sample; no causal estimate. A nonsignificant result is not proof that critical thinking cannot improve.

### WHAT LEADERS CAN DO

Ask the curriculum team to give each AI pilot two separate measures: students' confidence and an unaided reasoning task. Require a first attempt, a critique of the AI response, and an explanation of the revision. Review a sample of actual conversations before expanding a program described as Socratic tutoring.

### WHO / WHAT WAS STUDIED

Undergraduate applied statistics, spring 2025; indirect evidence for K-12.

### READ THE ORIGINAL SOURCES

**Thinking-partner pilot: full manuscript**
arxiv.org · Published 2026-09-29 · Open source

02 / LEADERSHIP

# Teach the concept through the model

Journal study / quasi-experiment
Barghot and Abu Elba / JETOL

### THE FINDING

A newly published study used machine learning as part of a hands-on electronics lesson sequence, rather than as a conversational tutor. Eighty eighth-grade boys in two existing classes studied the same sensing and remote-control unit. One class trained and tested an image classifier in Teachable Machine, then used PictoBlox to connect its classifications to electronic concepts. After 12 lessons, that class averaged 24.2 out of 25 on a curriculum-aligned test, versus about 7.75 in the conventional class. The reported difference is unusually large, and the intervention class nearly reached the test ceiling. The result supports investigating the complete teaching sequence; it does not isolate the contribution of machine learning from examples, practice and teacher guidance.

### THE EVIDENCE

Two classes of 40, studied in 2021, with a researcher-developed multiple-choice test. Students compared labeled images, examined classification errors and connected visual features to definitions. Individuals were not randomly assigned.

### LIMITS TO KEEP IN VIEW

One school, boys only, immediate assessment and no delayed transfer test. Class effects and the bundled teaching changes limit attribution.

### WHAT LEADERS CAN DO

Have a technology teacher adapt a small classification activity around an existing curricular concept. Judge whether students can explain unfamiliar examples without the tool. Use an independently reviewed assessment and a later check before investing in expansion; an exceptional single-site score difference is a reason to replicate carefully.

### WHO / WHAT WAS STUDIED

Male grade-eight students, North Gaza, 2021-22; newly published September 2026.

### READ THE ORIGINAL SOURCES

**Teachable Machine electronics study**
dergipark.org.tr · Published 2026-09-30 · Open source

03 / LEADERSHIP

# When attention controls obstruct learning

Preprint / mixed-methods field study
Meenakshi V. and colleagues / arXiv

### THE FINDING

Webcam monitoring can change the experience of learning even when it is intended to support concentration. Researchers examined a monitored online internship-course sequence using open-ended surveys and platform records. Eighteen of 51 survey respondents raised focus monitoring without being explicitly asked about it. Four of the 15 dropout respondents identified it as their main reason for leaving. Learners described ordinary movements triggering interruptions, repeated video segments and restrictions they found costly. Other respondents valued the structure, and course-wide survey averages were above the scale midpoint. The evidence therefore identifies specific sources of friction, not a general verdict against monitoring. Crucially, a platform flag records a detection event, not a verified lapse in attention.

### THE EVIDENCE

The preprint combines 15 dropout and 36 persisting-learner responses with separately collected telemetry: 14,529 flags from 448 flagged students, plus other feedback channels. The survey and telemetry populations cannot be matched.

### LIMITS TO KEEP IN VIEW

Voluntary, small qualitative sample; no unmonitored comparison. Neither the false-positive rate nor monitoring-caused dropout across all enrollees is established.

### WHAT LEADERS CAN DO

Before adopting an attention product, ask student services and IT to test ordinary movements and accommodation needs. Preserve work after an interruption, provide an accessible alternative and a human appeal, and measure time lost. Evaluate the consequence of a flag as carefully as the detector itself.

### WHO / WHAT WAS STUDIED

Online summer internship coursework in India; indirect evidence for school settings.

### READ THE ORIGINAL SOURCES

**Attention monitoring: full manuscript**
arxiv.org · Published 2026-09-30 · Open source

04 / LEADERSHIP

# A rubric changes the AI grade

Journal study / model evaluation
Ozdemir and Asci / JETOL

### THE FINDING

What happens when an AI grader receives more than a generic instruction to assign a score? A Turkish history study compared seven models and five prompting approaches against an instructor's ratings of 24 university students' answers. The same answers produced different levels of alignment as prompts changed. For example, the correlation between ChatGPT o3 and instructor scores rose from .56 with basic instructions to .76 with weighted criteria. But stronger correlation did not remove statistically significant score differences for that model. Researchers also observed fluent, lengthy answers receiving generous treatment despite incomplete content. The useful lesson is about assessment design: explicit criteria can improve some comparisons, while matching a ranking remains different from awarding the right score.

### THE EVIDENCE

Exploratory analysis of 72 responses to three questions. Statistical comparisons use 24 student composite scores, not 2,520 independent students or scoring events. The human reference was one instructor.

### LIMITS TO KEEP IN VIEW

Single language, course and small sample; model versions age quickly. Nonsignificant differences do not establish grading equivalence or fairness.

### WHAT LEADERS CAN DO

Require a local, teacher-scored validation set before any grading pilot. Compare absolute score errors and consequential pass/fail disagreements, including terse but correct answers. Keep a teacher responsible for the final grade and give students an understandable route to challenge it. Recheck performance whenever the model or rubric changes.

### WHO / WHAT WAS STUDIED

University history answers in Turkish; indirect evidence for K-12 assessment.

### READ THE ORIGINAL SOURCES

**Prompt design and history scoring**
dergipark.org.tr · Published 2026-09-30 · Open source

05 / LEADERSHIP

# Use predictions to open doors

Preprint / predictive observational study
Tschisgale and Neumann / arXiv

### THE FINDING

Could a prediction model identify students missing a worthwhile STEM opportunity? Researchers trained models on 282 German Physics Olympiad participants, then estimated first-round success potential for 1,103 nonparticipants in academic-track secondary schools. At the selected threshold, 43 nonparticipants were flagged as potentially successful. Thirty-two had already entered another selective STEM competition; eleven had no such prior participation. The finding complicates the idea that every missing entrant represents undiscovered talent: many were already engaged elsewhere. Prediction quality was modest, with a cross-validated area under the curve of .605 for the chosen model. These are estimates of unobserved outcomes, not proof that the selected students would advance or that others lack potential.

### THE EVIDENCE

The preprint compared three classifiers using repeated nested cross-validation and 31 indicators. A regularized regression model was selected. The threshold was calibrated using participants and then transferred to nonparticipants.

### LIMITS TO KEEP IN VIEW

Selection and population differences may weaken that transfer. Threshold choice changes the identified group; actual competition outcomes for nonparticipants are unavailable.

### WHAT LEADERS CAN DO

Ask enrichment leaders to examine who receives invitations, transport and preparation, including students without prior competition experience. Use any analytics to offer additional opportunities, with teacher and student input. Do not let an uncertain score become an eligibility barrier; track uptake and success after outreach.

### WHO / WHAT WAS STUDIED

German academic-track secondary schools; competition access, not U.S. district validation.

### READ THE ORIGINAL SOURCES

**Physics Olympiad opportunity analysis**
arxiv.org · Published 2026-09-28 · Open source

06 / LEADERSHIP

# Verified proof, imperfect interpretation

Preprint / classroom interface study
Guo and colleagues / arXiv

**THE FINDING**

A mathematically verified answer can still misrepresent what a student meant. LeanSide translates students' natural-language proofs into the Lean proof system, checks them, and returns readable feedback. Researchers studied 23 usable sessions alongside an advanced linear-algebra course deployment. Students sometimes repaired errors or made their reasoning more precise, but they also struggled with feedback that was technically correct and difficult to understand. Expert review found that the system changed the meaning of 22 of 117 incorrect steps, compared with 18 of 1,038 correct steps. Some changes silently repaired a student's error. Verification therefore helps only if the translation preserves the reasoning being assessed; a green indicator alone cannot demonstrate student understanding.

**THE EVIDENCE**

This preprint reports an interface study with expert judgments of reasoning steps and qualitative analysis of student revisions. The course deployment spanned 13 weeks; participants volunteered for study sessions.

**LIMITS TO KEEP IN VIEW**

No controlled learning-effect estimate. Advanced university mathematics and prototype behavior do not establish readiness for school mathematics or automated grading.

**WHAT LEADERS CAN DO**

In an AI feedback pilot, show the student's original reasoning beside the system's interpretation. Ask teachers to check whether a tool explains an error or silently solves it. Include incorrect and ambiguous work in acceptance testing, and retain a route to human help when verified feedback is unintelligible.

**WHO / WHAT WAS STUDIED**

University linear algebra; 21 of 23 analyzed participants were undergraduates.

**READ THE ORIGINAL SOURCES**

LeanSide: full manuscript
arxiv.org · Published 2026-09-30 · Open source

07 / LEADERSHIP

# Test what remains after AI help

Journal article / systematic review
Malikova and colleagues / Frontiers

### THE FINDING

A systematic review of university physics research found a varied picture behind broad claims about AI-supported learning. Eleven empirical studies met its criteria, covering tutoring, feedback, simulations, collaboration and teacher preparation. Some reported better immediate or task-specific outcomes. Others showed that apparent advantages did not persist on later examinations, depended on close alignment between homework and test questions, or coexisted with difficulty recognizing inaccurate AI explanations. The authors did not calculate a single pooled effect because designs and outcomes differed substantially. For school leaders, the transferable question is whether students can use what they learned independently and later. Successful completion of an AI-supported activity is only one part of that evaluation.

### THE EVIDENCE

The journal review searched six sources and used two reviewers for methodological appraisal. The included designs comprised two randomized studies, four nonrandomized studies, two descriptive studies and three mixed-methods studies.

### LIMITS TO KEEP IN VIEW

University-only evidence with uneven geographic coverage. Only the first 300 relevance-ranked Google Scholar results were screened; short-term and heterogeneous designs constrain conclusions.

### WHAT LEADERS CAN DO

For a science pilot, agree on the later assessment before launching. Include a new problem that requires transfer, not a repeat of the supported task. Have subject specialists check generated explanations and record misconceptions. Compare teacher preparation time with any correction burden as well as student performance.

### WHO / WHAT WAS STUDIED

University physics and preservice physics teachers across six countries; indirect K-12 evidence.

### READ THE ORIGINAL SOURCES

AI in university physics: systematic review
www.frontiersin.org · Published 2026-10-01 · Open source

08 / LEADERSHIP

# Ask students to demonstrate mastery

Preprint / retrospective assessment study
Sharma and colleagues / arXiv

**THE FINDING**

As schools reconsider assessment in an AI-rich environment, a programming-course study offers a useful design example rather than a test of AI itself. Researchers compared 203 students' reports of completed tasks with oral demonstrations across a 60-task sequence. Of 11,093 reported tasks, 10,885 were ultimately demonstrated: 98.1%. The remaining 208 mismatches clustered in the hardest late tasks and among a minority of students. However, students could try again, and the records retained final outcomes rather than attempt counts. The high agreement therefore describes eventual demonstration within this system, not accurate first-attempt self-assessment. It suggests where targeted conversations could be useful while leaving the staffing and learning benefits of that approach untested.

**THE EVIDENCE**

The preprint analyzes 2023 course records and 12 examiners. A task was examined only after a student reported it. This one-sided design measures unresolved overclaiming, not unrecognized competence.

**LIMITS TO KEEP IN VIEW**

One university course; examiner and student groups were not independently crossed. No AI intervention, cheating-detection test or comparison of assessment costs.

**WHAT LEADERS CAN DO**

Pilot brief explanations or demonstrations at a few consequential milestones. Use a common rubric, accessible response options and moderation between teachers. Treat the conversation as evidence of understanding and a chance for feedback, rather than assuming an unexplained mismatch proves misconduct. Record the staff time needed.

**WHO / WHAT WAS STUDIED**

First-year engineering programming, 203 students; assessment design evidence adjacent to AI.

**READ THE ORIGINAL SOURCES**

Self-assessment and oral mastery study
arxiv.org · Published 2026-09-30 · Open source

FROM EVIDENCE TO ACTION

# The week ahead

**01** Curriculum and assessment: Choose one current AI pilot and add an unaided reasoning task plus a delayed transfer question. Keep confidence ratings separate from achievement.

**02** Student services and IT: Review what happens when monitoring flags a student. Preserve completed work, test accommodations and provide human review.

**03** Enrichment leaders: Audit invitations, preparation and access to STEM opportunities. Evaluate outreach results rather than using a prediction as a gate.

**04** Superintendent cabinet: Ask each pilot owner to bring one observed benefit, one failure case, the staff cost and the evidence needed before expansion.

**KEEP READING**

Visit The Sethi Ledger · Explore the archive