

AI Tutors Meet the Learning Test: Access Alone Still Isn't an Intervention

AI in EdTech Weekly

2026-09-14

AI Tutors Meet the Learning Test: Access Alone Still Isn't an Intervention

By AI in EdTech Weekly • September 14, 2026

An independent two-year Tennessee trial found no measurable added math gain from Khanmigo access, while PISA's global decline and new age-tiered policies sharpen the distinction between technology exposure and learning.

The clearest development this week is a correction to the “just add a tutor” model: giving students access to an AI tutor is not the same as creating a learning intervention. In a Digital Learning Week discussion, the moderator described an independent two-year randomized trial across 18 Tennessee middle schools involving 11- to 14-year-olds in daily remedial math. Both groups improved somewhat, but the students with Khanmigo did not improve measurably more after two years. Sal Khan said the reported second-year effect size was 0.08, or about 0.14 after adjusting for students who left the intervention when they stopped struggling; the researchers said that adjusted result was consistent with Khan Academy's non-AI results. [1]

Khan's explanation is about use, not model quality alone. He said a small group of students used the tutor constructively, while most expected direct answers and became frustrated when its Socratic approach refused to provide them. Khanmigo's newer approach appears after a wrong answer and offers credit if students review the problem and demonstrate understanding; Khan described early engagement pilots as promising but said another year is needed to determine whether the AI adds efficacy. [1]

The practical shift is from optional access to structured engagement, teacher integration, and accountability. Khan said the strongest motivational forces are teachers and parents, not game mechanics. Wayne Holmes argued that ministries should demand evidence not only of test-score effects, but also of consequences for cognition, mental health, individuality, and classroom relationships.



[1]
Sal Khan & Wayne Holmes discussing AI Tutors (2:42)

From usage to evidence

A meta-analysis of Penda Learning across 13 independent school-based samples involving more than 33,000 students found significant science-achievement gains, but the strongest predictor was mastery rather than time on the platform. Schools with low, moderate, and high levels of mastered activities showed effect sizes of 0.56, 0.89, and 1.40; time-on-task alone was not positively associated with better performance. [2] The implication is broader than one science intervention: districts should define success in terms of mastery, retention, application, and teacher use of learning data—not logins, minutes, or completed activities. [2]

Procurement practice is not yet aligned with that standard. District interviews found that AI quality checks emphasize ethical-use frameworks, containment, hallucination checks, technical integration, and switching costs—mostly risk containment and operational fit rather than instructional efficacy. Only 40% of districts formally establish evaluation criteria before reviewing products, while 94% rely on peer recommendations and formal evidence intermediaries account for just 8% of market awareness. [3]

The implementation gap is visible in classrooms: 62% of teachers report that AI guidance is nonexistent, vague, or unclear, and 83% say managing AI is their largest daily challenge. Although 56% say AI should be teacher-led or

process-focused, more than half of students report using it to brainstorm and draft writing rather than engage deeply with learning. Districts are also warned to monitor products after adoption because models can change while sunk-cost bias makes reevaluation unlikely. [3]

PISA puts the screen-and-AI debate in sharper context

The OECD’s PISA 2025 release reports the lowest average OECD performance yet recorded in reading and mathematics: across 760,000 students in 91 countries and economies, reading scores fell 28 points from 2015 to 2025—about one and a half years of learning—and mathematics fell 22 points. One in five 15-year-olds was classified as a low performer in science, mathematics, and reading, up from 16% in 2022. [4]

The release does not offer a single technology explanation, but its comparisons are consequential. Students tended to perform less well when digital use became excessive or devices were used for leisure at school; more than one in four said classmates were distracted by devices in most or every science lesson. Students who said they did not use AI to draft writing outperformed those who said they did, while frequent users who were taught to assess AI-generated information tended to perform better than frequent users without that guidance. The wording points to an association—and to the importance of instruction about AI use—rather than a simple ban-or-adopt conclusion. [4]

Schools are drawing an age-based boundary around AI

New York City has now put its yearlong moratorium on student-facing generative AI in preschool through eighth-grade classrooms into operation. The policy bans individual devices in preschool through second grade and recommends no more than 30 minutes of individual-device use in grades 3–5 and 45 minutes in middle school. Parents told Chalkbeat that school-set limits make screen negotiations at home easier, but the guidance does not address AI use for homework, and the screen limits are guidelines rather than strict caps. [5]

Norway offers a more explicit developmental sequence. Its non-binding national guidance generally rejects generative AI in grades 1–7, permits cautious teacher-guided use in lower secondary, and emphasizes appropriate, critical, and responsible use in upper secondary. The rationale is that AI can displace the cognitive work through which younger learners develop language, memory, reasoning, and attention. The proposed governance model also calls for safeguards, data protection, professional capacity, early teacher and researcher participation, and decisions that can be revised as evidence and technology change. [6]

The low-tech correction is also colliding with budgets and infrastructure. Memphis-Shelby County Schools’ White Station High School opened without the one-to-one laptops or tablets students had received in prior years; the district said it was evaluating whether the traditional model remained the most

effective use of resources amid rising equipment and repair costs. The school mostly returned to paper, but shortages of printed materials and devices forced some students to bring personal devices for online tests—showing that reducing technology without a transition plan can create a different access problem. [7]

Early childhood has a content-quality problem, too

EdSurge reports that the internet now contains hundreds of thousands of AI-generated videos aimed at preschoolers and kindergartners. Educators describe much of this “baby slop” as low-substance, while examples include inappropriate scenes, inaccurate visuals, and channels producing thousands of videos in a short period. Researchers and teachers warn that repeated exposure can contribute to cognitive overload and make it harder for young children to distinguish reality from fiction. [8]

The response being recommended is active media literacy, not passive filtering: preview videos, inspect creators and channels, disable autoplay, ask children what they noticed, and compare an AI-generated image with photographs of the real thing. Parent follow-up questions can turn a video about numbers or community helpers into an opportunity to explain, demonstrate, and connect the content to the child’s world. [8]

Higher education and workforce learning are negotiating judgment

At the University of Sydney, academic staff included AI consultation guarantees in enterprise negotiations, including a commitment that changes forcing staff to use AI would not be imposed unilaterally. An Adelaide University agreement reportedly includes information-sharing, training, consultation, and no mandatory AI requirements. [9] At Macquarie University, a headline suggesting an AI chatbot was teaching online psychology classes prompted a clarification: students still use course materials, online modules, academic staff, and tutors, while the chatbot is an optional supplementary question-answering tool that does not replace teaching. [9]

EDUCAUSE’s focus groups with 30 online-learning leaders show why that distinction matters. Institutions are experimenting quickly, sometimes through decentralized, bottom-up work, but evidence of impact and return on investment is lagging because baseline data were not collected before implementation. The recommendation is not to stop experimenting; it is to define the purpose, collect baseline measures, evaluate the intervention, and avoid scaling simply because a pilot has already consumed time and money. [10]

A concrete workforce-learning model came from Gauntlet’s Cohort 6 showcase. Organizers described a 10-week, roughly 1,000-hour program built around hiring-partner projects followed by learner-chosen capstones; participants reported that domain-specific projects and peer immersion changed how they used AI and

prepared them for interviews. Those are program and participant accounts, not independent outcome evidence, but they illustrate a move toward demonstrating capability through realistic work rather than course completion alone. [11]

One capstone, Debate Coach AI, turns debate video, audio, or text into argument maps linked to evidence, identifies which arguments remain decisive, and produces a verdict with advice for the next round. Andrew Ng's current AI-engineering skills map describes the corresponding human capability: shaping specifications, driving rapid build-and-feedback loops, making product decisions, communicating across functions, and taking end-to-end ownership of the result. [11, 12]



Gauntlet AI Cohort 6 Showcase: 4 AI Capstone Demos in 10 Weeks (50:46)

What This Means

- **For AI-tutor buyers:** Treat a tutor as a learning intervention, not an account provision. Define the target skill, expected student effort, baseline, mastery measure, and unaided transfer check; require teacher integration and post-adoption monitoring. The Penda findings, Khanmigo trial, and higher-ed evidence gap all argue against treating usage as impact. [2, 1, 10]
- **For teachers and assessment leaders:** Preserve evidence of thinking before and during AI use. A practical RAG workflow retrieves teacher-selected sources, gives students an evidence trail, and still requires verifi-

cation because retrieval does not eliminate hallucinations. The classroom rule proposed this week is simple: never cite the AI tool; cite and inspect the evidence behind it. [13]

- **For K–8 leaders:** Age-based restrictions are only the first design decision. Publish what applies to homework, define who owns implementation, involve teachers and researchers, and budget for the paper, device, accessibility, and family-communication consequences of the transition. [5, 6, 7]
- **For higher education and L&D:** Pair AI literacy with consultation rights, domain fundamentals, applied projects, and judgment about what to build and whether it works. Employer-linked project work and Ng’s skills map point toward assessment through specifications, iteration, communication, and verification rather than prompt fluency alone. [9, 11, 12]
- **For early-childhood educators and families:** Treat AI-generated media as something to preview and discuss, not as a neutral stream of entertainment. Disable autoplay and use real-world comparison and follow-up questions to keep adult conversation at the center. [8]

Watch This Space

- **Khanmigo’s proactive redesign:** Khan says another year is needed to determine whether prompting students after wrong answers produces measurable gains beyond Khan Academy’s non-AI effects. [1]
- **Policy implementation:** Watch whether New York City fills the homework and enforcement gaps in its guidance, and whether Memphis can restore reliable access to print and online testing while reducing device dependence. [5, 7]
- **Evidence infrastructure:** Districts will need to turn formal efficacy criteria, baseline data, student voice, and dynamic post-adoption monitoring from recommendations into procurement practice. [3, 10]
- **Applied AI learning:** Employer-linked programs and products such as Debate Coach AI are promising demonstrations of practice-centered learning; the next question is whether they produce independently verified skill gains and durable workplace performance. [11, 12]

Sources

1. Sal Khan & Wayne Holmes discussing AI Tutors
2. Your Edtech Is Getting Used. But Is It Actually Working?
3. How AI is Reshaping K-12 EdTech Procurement
4. PISA 2025: Students’ reading and mathematics performance declined sharply across the OECD

5. NYC back to school: New AI and screen time limits give parents relief
6. Yes, No, Not Yet for Generative AI in Schools
7. Memphis schools stopped giving laptops to every student this year
8. AI Videos Are Warping Early Learning, Teachers Say
9. Is AI in Education Reaching an Inflection Point?
10. The Growing Influence of Online Learning in Higher Ed | Shop Talk
11. Gauntlet AI Cohort 6 Showcase: 4 AI Capstone Demos in 10 Weeks
12. X article by @AndrewYNg
13. Beyond ChatGPT: 5 Ways Teachers Can Use RAG to Help Students Find Better Information