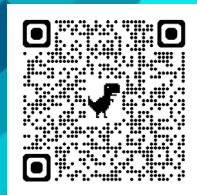


The Evidence on AI in K-12 Education

[And How Leaders Can Respond]

Chris Agnew | AI Hub for Education, Stanford University



**We're asking
education
leaders to
make
consequential
decisions about
AI at a moment
when...**

**We're asking
education
leaders to make
consequential
decisions about
AI at a moment
when...**

- **Technology is advancing rapidly**
- **Policy is still forming**
- **Evidence is still emerging**

**And most
districts are
currently
making AI
decisions in the
wrong order**

**And most
districts are
currently
making AI
decisions in the
wrong order**

**But first, let's
understand the
evidence**



The Evidence Base on AI in K-12: A 2026 Review

The existing research on the impacts
of AI on students and teachers



Lily Fesler JP Martinez Claeys Chris Agnew Susanna Loeb

Stanford | SCALE Initiative
Accelerator for Learning

- The current trends in AI research relevant to K-12
- Identifying the high-quality causal studies
- Distilling it down to key findings (and what's missing)

released
LAST MONTH



818

Academic Papers Analyzed

Sourced from the AI Hub Research Repository (Preprints & Journals as of Oct 2025).

1



Phase 1: AI Review

- Checking student/teacher populations
- Confirming RCT and QED tagging
- Screen for high potential papers based on What Works Clearinghouse (WWC) standards

2



Phase 2: Expert Human Methodology Screening

- Expert evaluation for rigorous causal evidence (RCTs and QEDs)
- Applied strict WWC standards
- Eliminating bias, ensuring baseline equivalence and control for confounding factors

3

20

High-Quality Causal Studies

4

Student Key Findings

Students as AI users

Teacher Key Findings

Teachers as AI users

Takeaway #1

**There is no
high-quality
causal research
yet on student-
facing AI in US
K-12
classrooms**



Takeaway #1

**There is no
high-quality
causal research
yet on student-
facing AI in US
K-12
classrooms**

Most rigorous studies occur in:

- higher education
- controlled settings
- international contexts

Takeaway #2

**Evidence on
student-facing
AI is mixed**



Takeaway #2

**Evidence on
student-facing
AI is mixed**

Outcomes likely
depend on
pedagogical design:

- step-by-step instruction → stronger effects
- open-ended AI tools → weaker or inconsistent results

Takeaway #3

**Evidence on
teacher-facing
AI is more
promising**



Takeaway #3

**Evidence on
teacher-facing
AI is more
promising**

Examples with positive results:

- lesson planning
- teacher feedback
- instructional support
- formative assessment support

Takeaway #4

**The evidence is
weakest where
stakes are
highest**



Takeaway #4

**The evidence is
weakest where
stakes are
highest**

Little to no high-quality research yet on:

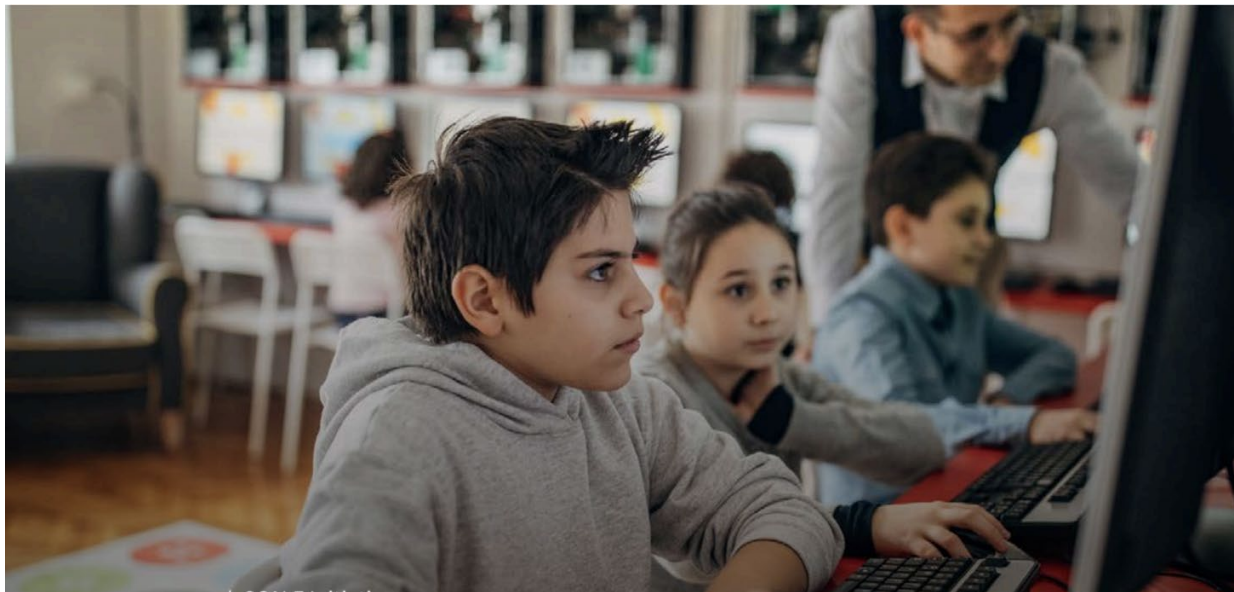
- equity impacts
- long-term learning
- social development
- student well-being

Education
leaders cannot
wait for perfect
evidence

Education
leaders cannot
wait for perfect
evidence

And they
cannot treat
early signals as
settled
conclusions

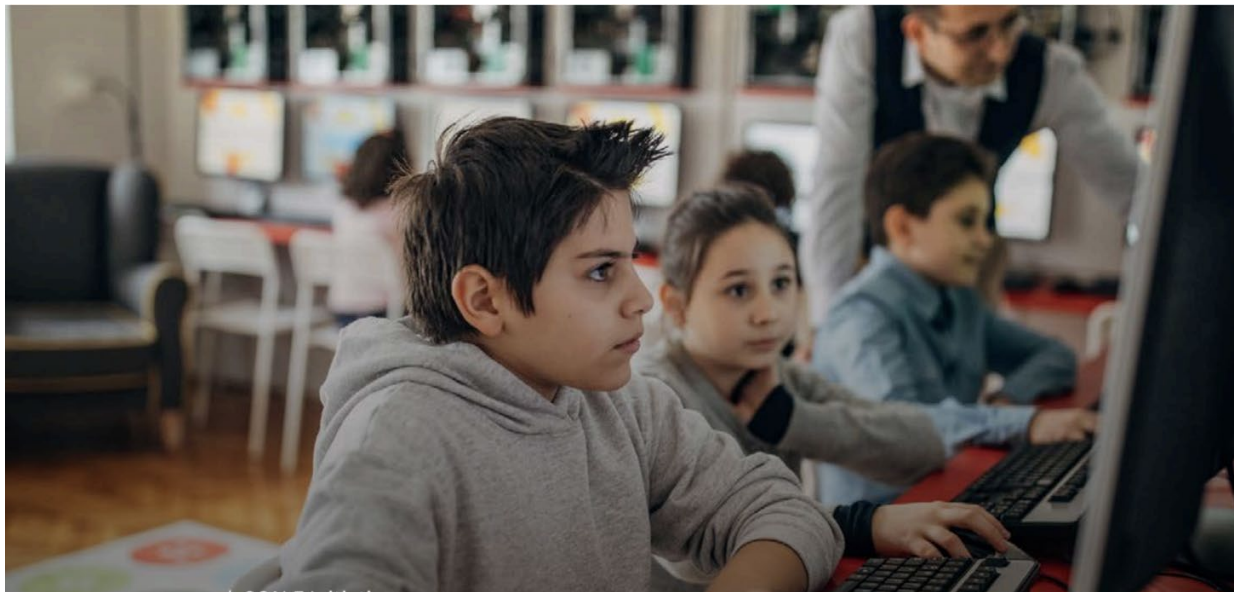
Most AI decisions start with tools



Common sequence:

1. AI tool →
2. Pilot →
3. Adoption

Most AI decisions start with tools



Common sequence:

1. AI tool →
2. Pilot →
3. Adoption

**But the state
of the
evidence
suggests this
is backwards**

AI in service of what?

AI in service of what?

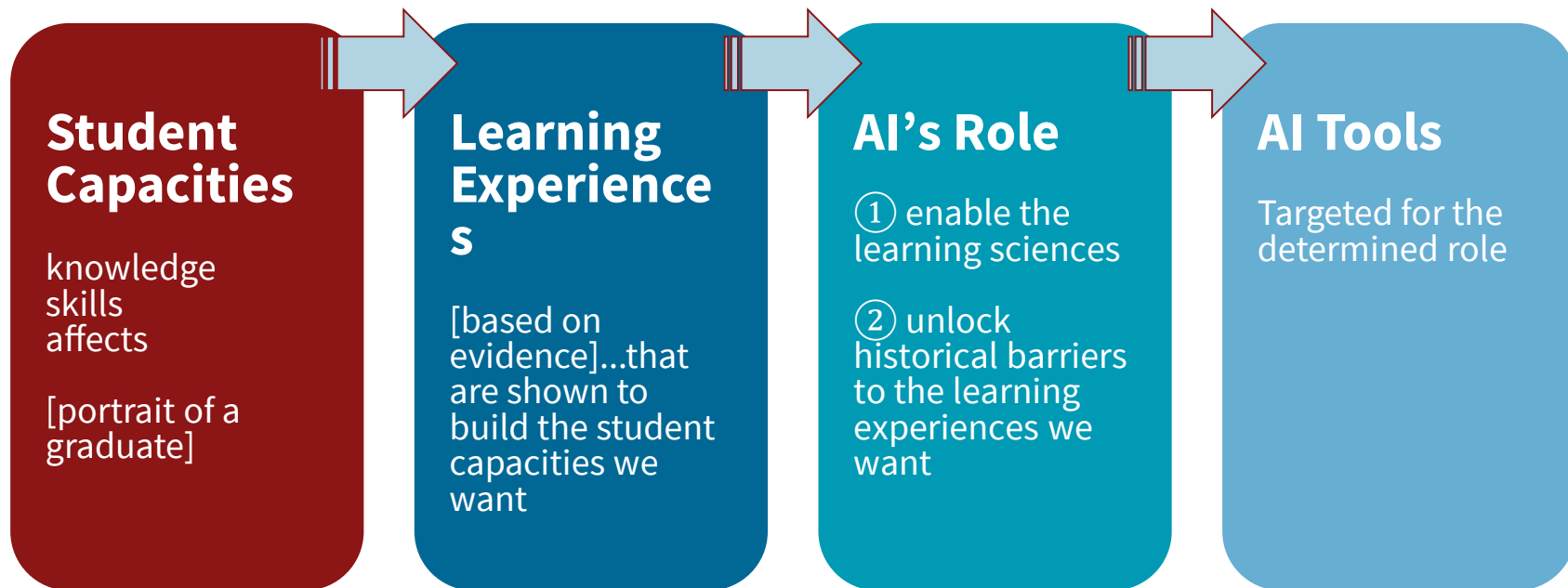
What we know to be true about learning science:

-  Cognitive Load Theory
-  Zone of Proximal Development
-  Transfer of Learning
-  Metacognition
-  Expertise Reversal Effect
-  Productive Struggle

Instead, start with student capacities



Instead, start with student capacities



Why Order Matters

Adopting tools first

- markets shape learning goals
- systems lock into products
- we optimize a legacy system with AI

Districts start with learning

- tools serve learning design
- innovation is aligned with what we want for students
- educators remain central

Given the current evidence...

- **Student-facing AI**
impacts are mixed [depends on pedagogical design]
- **Teaching-facing AI**
promise in improving instruction, feedback, formative assessment, & lesson planning
- **Evidence is lacking**
on equity, learning needs, long-term outcomes, and SEL
- **No causal studies**
on US K12 students



Given the current evidence...

- **Student-facing AI**
impacts are mixed [depends on pedagogical design]
- **Teaching-facing AI**
promise in improving instruction, feedback, formative assessment, & lesson planning
- **Evidence is lacking**
on equity, learning needs, long-term outcomes, and SEL
- **No causal studies**
on US K12 students

What does this mean for leadership decisions right now?

Table Discussion - 10 minutes

- **Student-facing AI**

impacts are mixed [depends on pedagogical design]

- **Teaching-facing AI**

promise in improving instruction, feedback, formative assessment, & lesson planning

- **Evidence is lacking**

on equity, learning needs, long-term outcomes, and SEL

- **No causal studies**

on US K12 students

1

Based on the evidence, **where specifically** do you want to be **cautious** on AI in YOUR DISTRICT?

2

Based on the evidence, where **where specifically** do you see the **most promise** on AI in YOUR DISTRICT?

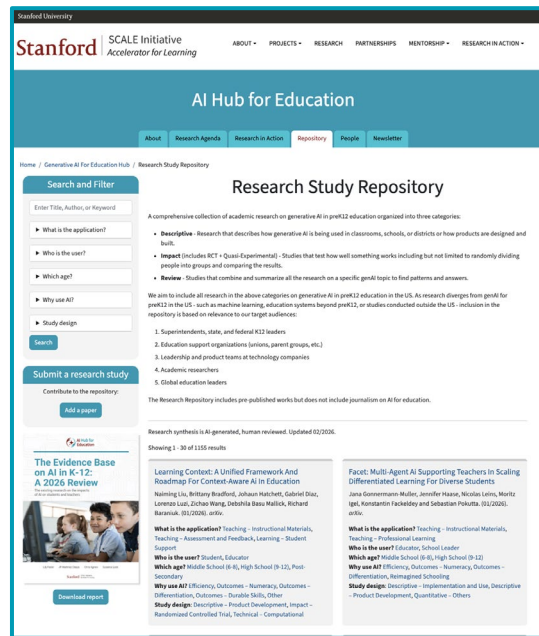
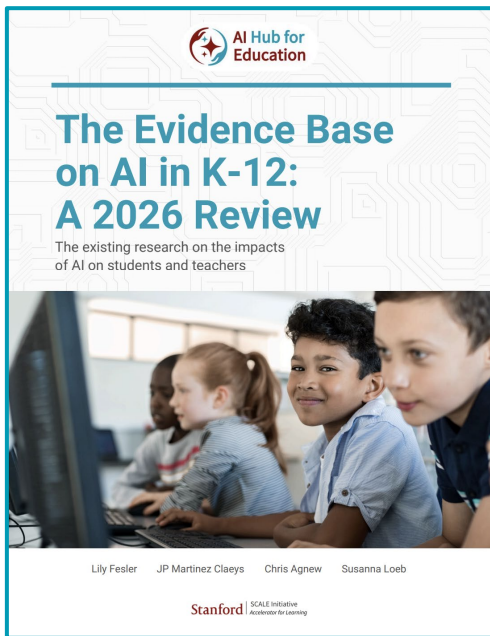
- Where might your district's expectations be ahead of evidence?
- What would your district want to *study before scaling* something?
- Where could your district accidentally move too fast?

What leaders can do now:

inaction
is not an
option

- 1 **Avoid a “tools-first” approach** - start with student capacities
- 2 Invest in **teacher-facing uses**
- 3 **Build AI literacy** - adults and students
- 4 **Avoid long-term vendor lock-in** too early
- 5 Stay connected to **evolving evidence**

Learn More



AI is reshaping education

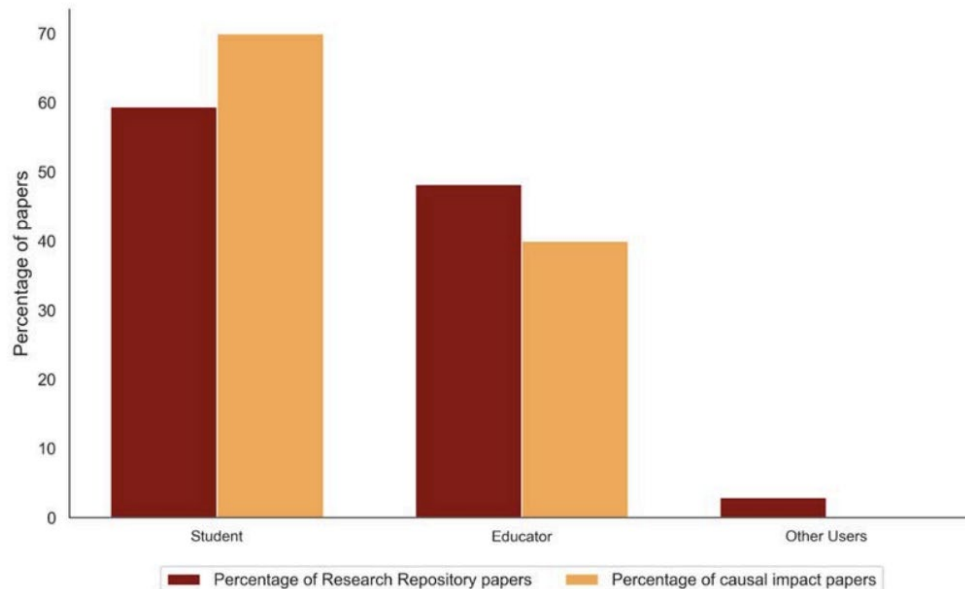
Actions now will determine if it:

- strengthens learning
- optimizes current flawed systems
- or erodes learning

Appendix

Research Trends

- Most AI in education papers focus on **students as users**.
- **Causal impact papers** disproportionately focus on the impacts of **AI on math skills**.
- **Most AI research** in education is conducted in **postsecondary settings**, but the **causal evidence** base is more **focused on high school** contexts.
- Only a **small proportion** of the research on AI in K-12 is **causal**.



Note: Studies can focus on multiple users, such as both students and educators. Other Users include school leaders and parents or caregivers.

20 High-Quality Causal Papers

Paper	User	Intervention Description	Sample	Key Finding					
Krejci et al. (2026) 	Students	Students studied two text passages using a GPT-3.5 chatbot, traditional note-taking, or a combination of both, following instructions for active, comprehension-focused reading.	344 Year 10 students (ages 14-15) in England	Note-taking (alone or with an AI chatbot) significantly outperformed AI-only use for comprehension and retention. Students preferred the AI chatbot and perceived it as more helpful, despite it leading to the lowest learning performance.	Degen et al. (2025) 	Students	Students interacted with a Socratic AI Tutor (scaffolded via structured dialogue) vs. an unstructured chatbot to refine a research question during a 5-minute session.	65 pre-service teacher students in Germany	The Socratic AI chatbot significantly increased perceived support for critical, independent, and reflective thinking compared to the unstructured AI chatbot.
Fischer et al. (2025) 	Students	Students used a GPT-4 based AI chatbot (with Retrieved Augmented Generation) to study economics material, comparing restricted access (10min delay) vs. unrestricted access vs. no access (textbook only) in a 25min session.	334 university students in Germany	Students with access to the AI chatbot outperformed students with no access to AI, and students with unrestricted access outperformed students with restricted access.	Meyer et al. (2024) 	Students	Students received AI-generated feedback (GPT-3.5-turbo) on an argumentative essay vs. no AI access. Conducted during a 90-minute classroom lesson.	459 upper secondary students (Grade 10) in Germany	AI-generated feedback significantly increased students' essay revisions, task motivation, and positive emotions compared to not receiving feedback.
Bastani et al. (2025) 	Students	Students engaged in GPT-4 math practice with pedagogical guardrails (GPT Tutor) vs. a standard interface (GPT Base). Students participated in four 90-minute sessions over one semester.	~1,000 high school students in Turkey	Unfettered AI access improved practice grades but led to a 17% performance drop on unassisted exams; guardrails (GPT Tutor) mitigated this "crutch" effect.	Zhao et al. (2025) 	Students	Students received AI autograder feedback for mathematical proof-writing vs. no AI access. Students completed one 60-minute assessment over the course of one week.	169 university students in the U.S.	Students using AI autograder feedback achieved significantly higher scores (approx. 11 points more) than those doing self-evaluation.
LearnLM Team, Google & Eedi (2025) 	Students, Educators	Students chatted with LearnLM (an AI tool fine-tuned for pedagogy and integrated into math tutoring chats and supervised by expert tutors) vs. chatted with human tutors. The intervention was over seven consecutive weeks.	165 Year 9/10 (ages 13-15) students across 5 schools in the UK	Students with AI-guided tutors performed at least as well as those with human tutors and showed superior knowledge transfer to novel topics.	Stadler et al. (2024) 	Students	Using ChatGPT-3.5 versus Google search to research scientific issues in a one-time 20-minute session and provide recommendations.	91 university students in Germany	AI chatbots significantly reduced students' cognitive load while they conducted research and reduced the quality of students' reasoning and argumentation compared to using a traditional search engine.
Kosmyrna et al. (2025) 	Students	Participants wrote essays using AI chatbot assistance (ChatGPT), search engines, or no tools. A subsequent condition swap tested the effects of removing AI support.	55 undergraduates, postgraduates, and university staff.	Writing essays using an AI chatbot reduced recall, with 83% of participants failing to be able to provide a quote from their essay, compared to 11% of participants who used a search engine or no tools.	Ferman et al. (2021) 	Students, Educators	Students received writing feedback from only AI vs. AI and independent human graders vs. not receiving any writing feedback. The program was implemented over one academic year.	Approximately 19,000 public school students in Brazil across 178 schools	Receiving AI feedback significantly improved student essay scores compared to the control group; the addition of human graders did not further improve performance despite students perceiving the feedback to be higher quality.
Blasco & Charisi (2025) 	Students	Students interacted with an AI chatbot that gave step by step reasoning vs. an AI chatbot that gave the solution alone, and a Socratic AI chatbot vs. a direct AI chatbot. Users completed one 60-minute session with the AI chatbot.	122 high school students in Belgium and Spain	The AI chatbot that provided step-by-step reasoning significantly improved performance compared to receiving the solution. The Socratic chatbot increased engagement but was perceived as less helpful than the direct chatbot.	Demszky et al. (2025) 	Educators	TeachFX delivered automated feedback to teachers via a mobile app and weekly emails that targeted "focusing questions" (to probe student thinking) vs. no feedback. The intervention took place in-person over five months.	369 math and science teachers in the U.S.	Automated feedback increased teachers' use of focusing questions by 20%. There was no discernible effect on student talk time or reasoning.
Becker et al. (2025) 	Students	Students used a custom-configured AI chatbot vs. tiered hints vs. a digital textbook page to solve a physics problem.	273 9th grade students in Germany	Both the AI chatbot and tiered hints significantly reduced cognitive load and improved affective outcomes (such as enjoyment and hope) compared to the digital textbook condition.	Wang et al. (2025) 	Educators	Tutor CoPilot gave real-time, expert-like AI suggestions to human tutors during math sessions vs. no suggestions. Students received tutoring over the course of two months.	900 tutors and 1,800 K-12 students from Title I schools in the U.S.	Access to Tutor CoPilot significantly improved student mastery of lesson topics (4-14 p.p.), and the benefits were largest for students of lower-rated and less-experienced tutors.
Chen et al. (2025) 	Students	Students had access to LLM-Tutor (which provided AI feedback on LaTeX proof submissions and a math-themed AI chatbot) vs. no AI feedback. It was a semester-long study.	155 university students in a discrete math course in the U.S.	Early AI access significantly improved homework performance, but had no significant impact on unassisted exam scores.	Roy et al. (2024) 	Educators	ChatGPT for lesson preparation (using a specific guide) vs. non-AI preparation. Teachers in the treatment group had access to ChatGPT for ten weeks and were provided with a guide on teaching with ChatGPT.	259 Year 7/8 science teachers across 68 schools in England	ChatGPT saved teachers an average of 25 minutes per week (31% reduction) on lesson preparation while maintaining the same quality.
Lehmann et al. (2025) 	Students	Students had unrestricted AI chatbot access (ChatGPT 3.5) while learning Python programming vs. no AI access.	Postsecondary students in Germany and the Netherlands.	AI chatbot access had no overall effect on learning; AI use increased the volume of topics covered but harmed student understanding. It also widened achievement gaps for students with low prior knowledge.	Demszky et al. (2023) 	Educators	M-Powering Teachers (automated NLP-based feedback to teachers on their "uptake" of student ideas) vs. no feedback. Access to the automated feedback was provided over 10 sessions.	414 instructors from a U.S.-based online platform for high school students	Automated feedback improved instructors' uptake of student ideas by 10%, reduced instructor's talk time by 5% and improved students' experience with the program.
					Kim et al. (2021) 	Educators	Providing AI-generated weekly student progress/achievement reports to home tutors vs. no reports. Tutors in the treatment group received AI-generated reports for 12 weeks.	234 tutors and 2,220 students in South Korea	AI reports significantly improved student academic performance, though "technology overload" hindered the benefit for certain tutors.
					Holstein et al. (2018) 	Educators	Access to mixed-reality smart glasses providing real-time AI-generated student analytics to teachers vs. no access to analytics. Students worked with a tutor for a total of 60 minutes, spread across two classes over one week.	286 middle school students and 8 teachers in the U.S.	Smart glasses allowed teachers to redirect attention to struggling students, narrowing the learning gap across prior performance levels.

Stanford

SCALE Initiative
Accelerator for Learning

The Learning Science Lens

Learning Principle	Description	AI Opportunity and Risk
Cognitive Load Theory (Sweller, 1988)	Manages the limited capacity of working memory by balancing intrinsic, extraneous, and germane (productive) loads.	AI can reduce extraneous load by efficiently retrieving and organizing information, potentially freeing cognitive resources for deeper learning, but it can also reduce germane load — the productive struggle essential for learning.
Vygotsky's Zone of Proximal Development (Vygotsky, 1978)	The optimal learning zone between what a learner can do independently and what they can achieve with appropriate support (Vygotsky, 1978).	The most effective AI tools would provide scaffolds within this zone and gradually release responsibility to the learner to prevent student dependency.
Transfer of learning	The process of applying knowledge gained in one context to new situations, which often requires explicit instructional support connecting the contexts (Barnett & Ceci, 2002).	One key question is whether practice with AI tools develops durable knowledge and skills that students can apply in new contexts, or whether it creates tool-dependent performance.
Metacognition	The ability of students to monitor their understanding, identify gaps, select appropriate strategies, and adjust their approach based on feedback.	Metacognition is difficult to measure and AI could measure metacognition at scale. At the same time, when AI tools perform complete tasks for students, opportunities to develop metacognitive skills may be reduced.
The expertise reversal effect (Kalyuga, 2007)	The phenomenon where instructional techniques effective for novices (such as worked examples) become ineffective or even counterproductive for more advanced learners (who may benefit more from independent problem solving).	Effective AI tools would adapt their support level to learner expertise.
Desirable difficulties (Bjork, 1994; Bjork & Bjork, 2011)	Challenges during practice that produce better long-term retention and transfer, even though they feel less effective and produce lower immediate performance (Bjork, 1994; Bjork & Bjork, 2011).	AI tools would ideally introduce appropriate desirable difficulties, even if users prefer easier practice sessions.

Key Findings - Students

- 14 causal studies
- 7 studied university students
- 7 studied high school students
- no US K-12 studies
- limited insights on equity considerations and learning needs, ELL, and students with disabilities
- no insights on the impacts on student emotional and social needs

1. Some student-facing AI tools **improve performance while students have access**, but effects **when students no longer have tool access are mixed**.
2. AI tools can **alleviate students' cognitive burden** and **foster positive experiences** in learning, although they **may not encourage deeper thinking**.
3. Tutoring chatbots may be **more effective than general-purpose AI chatbots**, and the **pedagogical design** of the tutoring chatbot **matters**.

Key Findings - Teachers

- 8 causal studies
- 4 studied K-12 in the US
- 4 studied K-12 in international contexts
- studies included teacher time use, teacher outputs, real-time feedback, tutoring, and types of teachers benefitting

1. AI can **reduce teacher time** spent on routine tasks, or shift teacher effort, with **no evidence of quality losses**.
2. AI can **deliver automated feedback** on teaching and student progress, **improving instructional quality** and **student outcomes**.
3. AI can offer **real-time instructional suggestions** that **improve educator practice** and **student outcomes**, particularly in messaging-based settings.
4. AI pedagogical supports appear to be **most beneficial to lower-rated** and **less experience educators**.