International Journal of Social Trends
Volume 4, Issue 6 (Nov-Dec; 2026)

Review Article · Education and the Learning Sciences

Can Machines Teach? A Review of AI Tutoring, Learning Analytics and Generative AI in Education

Natasha Kouassi [1], **Mwila Jean** [2]

[1] University of Botswana, BOTSWANA
[2] University of Namibia, NAMIBIA

ABSTRACT

Artificial intelligence has been promoted as a way to give every learner the benefits of a personal tutor. This review examines the evidence behind that promise across three generations of technology: intelligent tutoring systems, which model students' knowledge and guide them step by step through problems; learning analytics, which use data from digital learning environments to predict and support student success; and generative AI, whose conversational abilities have revived interest in automated tutoring. We organise the review around six questions that educators, researchers and policymakers commonly ask. The evidence suggests that well-designed intelligent tutoring systems produce meaningful learning gains, approaching those of human tutors in some comparisons, though effects on standardised tests are smaller. Learning analytics can identify students at risk but improve outcomes only when predictions lead to effective action. Early experimental evidence on generative AI shows that unrestricted access can raise performance during practice while harming learning, whereas tutors designed to guide rather than give answers can avoid this harm. We discuss equity, privacy and the role of teachers, and conclude that the value of AI in education depends less on technical capability than on pedagogical design.

Keywords: Intelligent Tutoring Systems; Learning Analytics; Educational Data Mining; Generative AI; Large Language Models; Effect Sizes; Educational Equity

INTRODUCTION

WHY ASK WHETHER MACHINES CAN TEACH?

In 1984 Benjamin Bloom reported that students tutored one-to-one performed about two standard deviations better than students taught in conventional classes, so that the average tutored student outperformed 98 percent of the comparison group [1]. He called the challenge of achieving such results with group instruction "the 2 sigma problem". Because one-to-one human tutoring is too expensive to provide to every student, the finding has inspired four decades of efforts to automate tutoring with computers.

Those efforts have passed through several stages. Intelligent tutoring systems (ITS), developed from the 1980s onward, model the knowledge a student is expected to acquire and give feedback on each step of a problem [2, 3]. In the 2010s, the growth of online learning produced large datasets of student activity and two related research fields, educational data mining and learning analytics, which aimed to use those data to understand and improve learning [4, 5]. Since late 2022, large language models have made fluent conversation with a machine widely available, renewing claims that AI tutors could transform education [6].

This review asks what the evidence says about each of these developments. Rather than surveying the field by technology, we organise it around questions that teachers, school leaders, researchers and policymakers ask. Each section ends with a short summary of key takeaways. We draw mainly on meta-analyses, randomised experiments and systematic reviews, and we report effect sizes where available, with the caution that effect sizes in education depend heavily on study design, outcome measures and populations [7].

How to read effect sizes in this review. Most studies report a standardised mean difference (Cohen's d or Hedges' g): the difference in average outcomes between groups divided by the standard deviation. An effect of 0.2 is often called small and 0.8 large, but these labels are misleading in education, where effects measured on broad standardised tests are typically much smaller than effects on tests designed by researchers to match what was taught [7]. When comparing results across studies, check what the comparison group received and how learning was measured.

WHAT DOES THIS REVIEW COVER?

We define AI in education broadly to include any system that uses computational models of learners, content or instruction to adapt teaching or to inform educational decisions. This includes intelligent tutoring systems, adaptive learning software, predictive learning analytics, automated assessment and feedback, and tools built on large language models. We exclude general-purpose technology, such as devices and connectivity, except where evidence on it helps interpret the effects of AI [8].

In selecting evidence we gave priority to meta-analyses and systematic reviews, which summarise many studies, and to randomised experiments, which provide the most credible estimates of causal effects. Where such evidence was lacking, as for many uses of generative AI, we draw on the strongest studies available and note their limitations. We concentrate on outcomes for learning, measured by tests and assessments, and give less attention to outcomes such as engagement, satisfaction or time on task, which are often reported but are imperfect indicators of learning [9]. Most of the evidence comes from mathematics and science in schools and universities in high-income countries; we note where findings come from other subjects and settings.

The review is structured as a sequence of questions. We begin with the history of educational technology and with Bloom's two-sigma benchmark, then examine evidence on intelligent tutoring, student modelling, learning analytics, automated feedback and generative AI. We then turn to equity and ethics, to the role of teachers and to methodological problems in the evidence base, before offering recommendations and a summary answer to the question in our title.

WHAT DID EARLIER GENERATIONS OF EDUCATIONAL TECHNOLOGY TEACH US?

Claims that machines will transform teaching are not new. In the 1950s B. F. Skinner proposed "teaching machines" that would present material in small steps, require an active response at each step and give immediate feedback, allowing each student to proceed at their own pace [10]. Many of these principles, including active responding, immediate feedback and self-pacing, reappear in modern intelligent tutors. Successive waves of technology since then, from film and television to personal computers and the internet,

have each been accompanied by predictions of revolution that were not fulfilled. A study of computers in American schools around 2000 found that despite heavy investment, most teachers used them rarely and in ways that left their teaching largely unchanged, a pattern described as "oversold and underused" [11].

Massive open online courses offered a more recent lesson. Launched with the promise of providing high-quality education to anyone with an internet connection, they attracted millions of learners, but completion rates remained low and the learners who benefited most were disproportionately those who already held degrees and lived in wealthy countries [12]. Reich's broader analysis of educational technology argued that new tools tend to be absorbed into existing practices rather than transforming them, and that technologies requiring self-directed learning often benefit already-advantaged learners most [13].

These histories do not show that technology cannot improve learning; the evidence on intelligent tutoring reviewed below shows that it can. They show that the effects of technology depend on how it is integrated into teaching, on who has the skills and support to use it, and on whether it embodies sound principles of learning. Advocates of AI in education have often underplayed these conditions, emphasising technical capabilities over pedagogical and institutional ones [14, 15]. The rest of this review should be read with these lessons in mind.

Key takeaways.

- Core tutoring principles, active response, immediate feedback and self-pacing, are decades old.
- Past technologies were often oversold and absorbed into existing practice.
- Self-directed technologies have tended to benefit already-advantaged learners.

DID THE 2 SIGMA PROMISE HOLD UP?

Bloom's two-sigma result has been influential, but later research suggests that it overstated the typical effect of human tutoring. In a widely cited review, VanLehn compared human tutoring, intelligent tutoring systems and other computer-based tutoring with no tutoring. He found that human tutoring raised outcomes by an average of about 0.79 standard deviations, far less than two sigma. More strikingly, step-based intelligent tutoring systems, which give feedback on each step of a solution, achieved an average effect of 0.76, nearly as large as human tutors. Systems that only check final answers achieved 0.31, and systems that operate at an even finer "substep" level achieved 0.40 (Figure 1) [16].

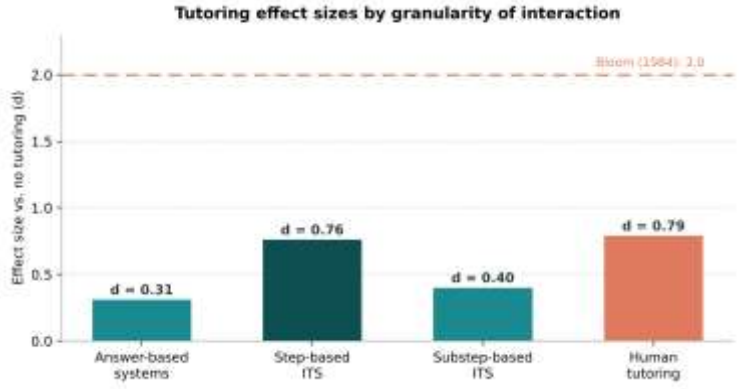


Figure 1: Average effect sizes of tutoring compared with no tutoring, by granularity of interaction, with Bloom's two-sigma claim shown for reference. Data sources: [16]; [1].

VanLehn's explanation for these findings was the "interaction granularity hypothesis": tutoring is effective largely because it gives students feedback and guidance at the level of individual reasoning steps. Moving from answer-level to step-level feedback brought large gains, but going finer still did not bring further improvement [16]. This suggests that the active ingredient in tutoring is not human presence as such, but timely, specific feedback on students' reasoning, which machines can provide.

The two-sigma figure likely reflected features of Bloom's studies, including mastery-learning procedures, the outcome measures used and the comparison conditions, rather than the typical effect of tutoring. Mastery learning, in which students must demonstrate mastery of each unit before moving on and receive corrective instruction when they do not, has positive effects of its own in meta-analyses [17], so the comparison in Bloom's studies combined two powerful interventions. This matters because it sets realistic expectations. A technology that matched average human tutoring would be a substantial achievement; expecting it to produce two-sigma gains sets a benchmark that human tutors themselves rarely meet.

Key takeaways.

- Human tutoring has a large but not two-sigma effect: about 0.79 SD in one prominent synthesis.
- Step-based intelligent tutors came close to human tutors in that synthesis.
- Feedback on individual reasoning steps appears to be the key ingredient.

HOW EFFECTIVE ARE INTELLIGENT TUTORING SYSTEMS?

What the meta-analyses show

Several meta-analyses have assessed ITS more broadly. Ma et al. synthesised 107 effect sizes and found that ITS produced higher achievement than every comparison condition they examined. The average advantage was 0.42 standard deviations over teacher-led, large-group instruction, 0.57 over non-ITS computer-based instruction and 0.35 over textbooks or workbooks [18]. Kulik and Fletcher reviewed 50 controlled evaluations and found a median effect of 0.66, but with a crucial qualification: effects were much smaller on standardised tests (0.13) than on tests developed locally by researchers (Figure 2) [19].

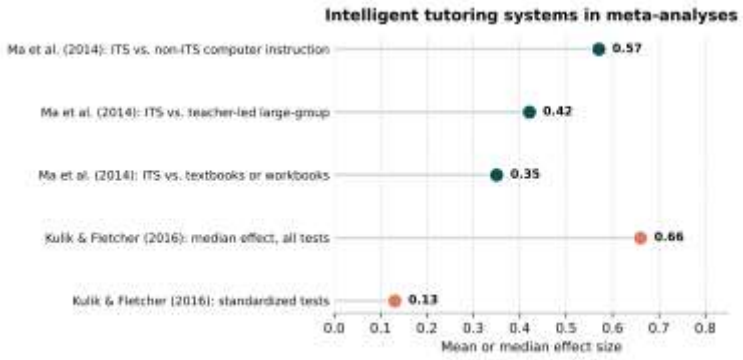


Figure 2: Effect sizes for intelligent tutoring systems reported in two meta-analyses. Data sources: [18]; [19].

Two further meta-analyses by Steenbergen-Hu and Cooper focused on particular populations. In K–12 mathematics, they found that ITS had small effects overall, and smaller effects in more recent and larger studies [20]. Among college students, the effects were moderate and positive, and ITS outperformed most other forms of instruction except individual human tutoring [21]. Table 1 summarises these syntheses.

Table 1: Meta-analyses and reviews of intelligent tutoring systems.

Synthesis	Scope	Main finding	Important qualification
VanLehn (2011)	Human, intelligent and other tutoring compared with no tutoring	Step-based ITS (0.76) nearly as effective as human tutoring (0.79)	Effects depend on granularity of feedback
Ma et al. (2014)	107 effect sizes across levels and subjects	ITS outperform teacher-led instruction (0.42), other computer instruction (0.57), textbooks (0.35)	Heterogeneity across settings and designs
Steenbergen-Hu & Cooper (2013)	K–12 mathematics	Small positive effects overall	Smaller effects in larger and more recent studies
Steenbergen-Hu & Cooper (2014)	College students	Moderate positive effects	Not better than one-to-one human tutoring
Kulik & Fletcher (2016)	50 controlled evaluations	Median effect 0.66	Only 0.13 on standardised tests

Note. Sources: [16], [18], [20], [21], [19].

Evidence at scale

Meta-analyses mostly summarise studies conducted by researchers, often with the developers of the system involved. Large-scale field trials give a more demanding test. The Cognitive Tutor Algebra I curriculum, which grew out of the research programme described by Koedinger and colleagues [2], was evaluated in a randomised trial across many schools in several US states. In the first year of implementation, the curriculum had no significant effect on high-school students' algebra scores; in the second year, a positive effect emerged [22]. The pattern suggests that schools and teachers need time to learn how to use such systems effectively.

Evidence from developing countries is also encouraging. A randomised evaluation of Mindspark, a computer-based adaptive learning program used in after-school centres in Delhi, found large gains in mathematics and Hindi over a few months, with particularly large benefits for students who were furthest

behind grade level [23]. A review of experimental research on education technology concluded that computer-assisted learning software, especially software that adapts to the student's level, is among the most consistently effective uses of technology in education, while simply providing hardware such as computers or tablets generally does not improve learning [8].

Other large evaluations point in the same direction. ASSISTments, a free online platform that gives students immediate feedback on mathematics homework and gives teachers reports on common errors, was evaluated in a randomised trial in schools in Maine. Students in schools using the platform scored higher on an end-of-year mathematics test, with the largest benefits for students with lower prior achievement [24]. The intervention is notable for its simplicity: it did not attempt sophisticated student modelling, but combined immediate feedback for students with actionable information for teachers.

Intelligent tutors have also moved beyond mathematics. AutoTutor, for example, holds conversations in natural language with students about topics in physics and computer literacy, asking questions, giving hints and correcting misconceptions in a manner modelled on human tutors [25]. Dialogue-based systems of this kind anticipated many features of today's language-model tutors, and the research on them is a useful guide to what such tutors must do to support learning.

Key takeaways.

- ITS produce meaningful gains over most alternatives in controlled studies.
- Effects on broad standardised tests are much smaller than on researcher-designed tests.
- Implementation matters: gains at scale may take time to emerge.
- Adaptive software helps; hardware alone does not.

HOW DO THESE SYSTEMS MODEL WHAT STUDENTS KNOW?

An intelligent tutor must estimate what a student knows in order to decide what to do next. The most influential approach, Bayesian knowledge tracing, models each skill as either learned or not learned and updates the probability that a student has learned it after each correct or incorrect response, taking account of the chances of guessing correctly or slipping [3]. These estimates allow a tutor to give students more practice on skills they have not mastered and to move on when mastery is likely.

Machine learning has extended this approach. Deep knowledge tracing uses recurrent neural networks to model sequences of student responses without specifying the skill structure in advance, and the original study reported improved prediction of student performance [26]. Educational data mining more broadly has developed methods for predicting performance, detecting disengagement, discovering the structure of knowledge in a domain and modelling students' affective states [5]. Table 2 summarises the main approaches.

Table 2: Approaches to modelling student knowledge and behaviour.

Approach	How it works	What it enables	Limitation
Bayesian knowledge tracing	Probabilistic model of mastery per skill, updated after each response	Mastery-based practice; interpretable estimates	Requires a predefined skill map
Deep knowledge tracing	Recurrent neural network over response sequences	Better prediction without skill labels	Less interpretable; gains debated
Predictive analytics	Classification models predicting course outcomes from activity data	Early warning of students at risk	Prediction does not ensure effective intervention
Behaviour and affect detection	Models of engagement, gaming the system, confusion	Responsive tutoring and teacher alerts	Validity across contexts uncertain
Language-model tutoring	Large language models generating dialogue and hints	Open-ended conversation in many subjects	Errors; may give answers rather than support learning

Note. Sources: [3], [26], [5], [27], [6].

Better prediction is not the same as better teaching. A model that predicts accurately whether a student will answer correctly is useful only if the tutor uses that prediction to choose instruction that helps the student learn. Much research on student modelling reports predictive accuracy; much less tests whether improved models lead to improved learning. This gap between prediction and pedagogy recurs throughout the field.

Key takeaways.

- Knowledge tracing lets tutors decide when a student has mastered a skill.
- Machine learning improves predictions, but improved prediction must be tested against improved learning.

DOES LEARNING ANALYTICS IMPROVE STUDENT SUCCESS?

From data to action

Learning analytics is commonly defined as the measurement, collection, analysis and reporting of data about learners and their contexts, for the purpose of understanding and improving learning and the environments in which it occurs [4, 28]. The field grew with the spread of learning management systems and online courses, which record detailed traces of student activity. Its central idea is captured by the learning analytics cycle: learners generate data, data are turned into metrics and predictions, and these inform interventions that change what learners do (Figure 3) [29].

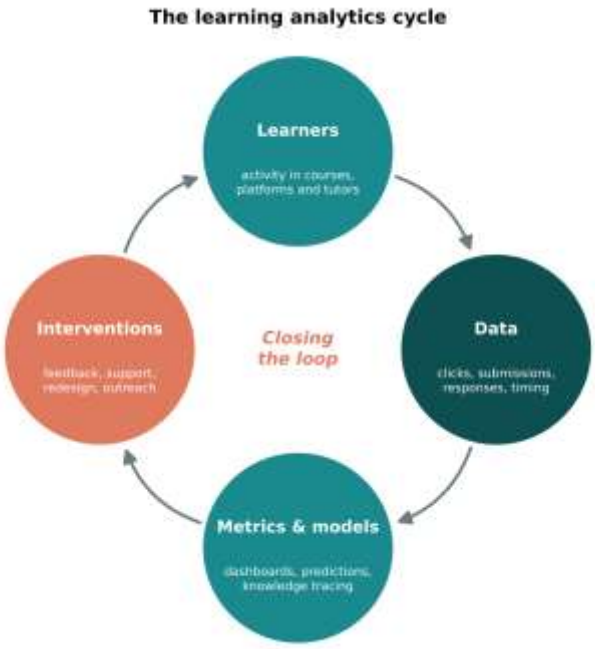


Figure 3: The learning analytics cycle. Source: authors' design, adapted from the description in [29].

The cycle emphasises that analytics creates value only when the loop is closed: when data lead to action that improves learning [29]. A dashboard that no one consults, or a prediction that triggers no support, changes nothing. Researchers have also warned against analytics that measure what is easy to record, such as clicks and logins, rather than what matters for learning, and have argued that the field must stay grounded in learning theory [9].

Early-warning systems

The best-known early application was Course Signals at Purdue University, which combined grades, demographic characteristics, academic history and activity in the learning management system to classify students into risk categories displayed as traffic-light signals. Instructors could then send messages and refer students to support. Purdue reported higher grades and retention among students in courses using the system [27]. The retention findings were later criticised on methodological grounds, because students who took more courses using Signals were also more likely to stay enrolled for reasons unrelated to the system. The case illustrates both the appeal of early-warning analytics and the need for rigorous evaluation.

A systematic review of AI applications in higher education found that most studies focused on profiling and prediction, intelligent tutoring, assessment and adaptive systems, and noted a striking lack of critical reflection on pedagogical and ethical issues and of involvement by educators in the research [15]. The review's subtitle, asking "where are the educators?", captured a concern that recurs across the field: technologies designed without teachers may predict outcomes well but fit poorly into teaching practice.

Key takeaways.

- Learning analytics can identify students who are likely to struggle.
- Benefits depend on effective action following prediction ("closing the loop").
- Rigorous causal evaluations of analytics interventions remain scarce.

WHAT HAPPENS WHEN STUDENTS USE GENERATIVE AI?

New capabilities, new risks

Large language models can explain concepts, generate practice problems, give feedback on writing and hold extended conversations on almost any topic. A review of their educational potential highlighted opportunities such as personalised explanations, support for teachers and assistance for students with disabilities, alongside risks including factual errors, bias, over-reliance and academic dishonesty [6]. UNESCO guidance has urged governments to regulate the use of generative AI in education, including setting age limits for independent use and protecting data privacy, and to build teachers' capacity to use it well [30].

Experimental evidence: practice is not learning

The most informative evidence so far comes from a randomised experiment with about 1,000 high-school students in Turkey [31]. Students in mathematics classes were assigned to practise with no AI support, with a chatbot based on GPT-4 that resembled a standard chat interface ("GPT Base"), or with a version designed with teachers' input to give hints and guidance rather than full answers ("GPT Tutor"). During practice, both AI versions improved performance substantially: by 48 percent for GPT Base and 127 percent for GPT Tutor relative to the control group. On a subsequent exam taken without AI access, however, students who had practised with GPT Base performed 17 percent worse than the control group, while those who had used GPT Tutor performed about the same as the control group (Figure 4) [31].

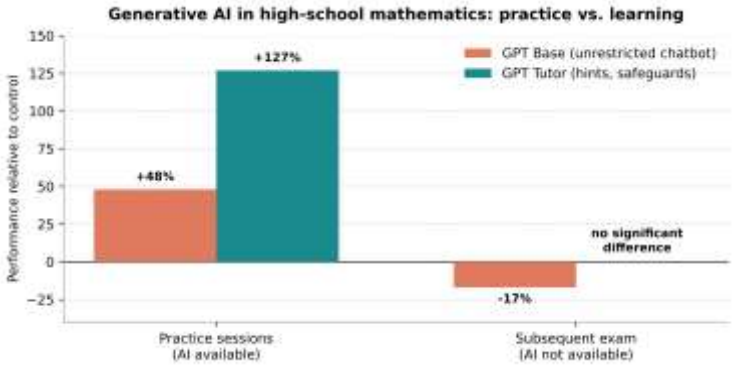


Figure 4: Effects of two versions of a GPT-4-based assistant on performance during practice and on a later exam without AI access, relative to a control group. Data source: [31].

The authors attributed the harm from GPT Base to students using the chatbot as a crutch, asking for answers rather than working through problems. This finding is consistent with long-standing research on learning. Learning requires active cognitive engagement; the ICAP framework distinguishes passive, active, constructive and interactive modes of engagement and finds that deeper modes produce more learning [32]. Research on "desirable difficulties" shows that conditions that make practice harder, such as retrieving

information rather than rereading it, often produce better long-term retention even though they reduce performance during practice [33]. An AI that removes difficulty by supplying answers can raise immediate performance while undermining the effort that produces learning.

The GPT Tutor result is equally important. When the AI was designed to support rather than replace students' thinking, the harm disappeared, although the experiment did not show a gain on the exam either [31]. This suggests that the pedagogical design of AI tools, including what they will and will not do for students, is as important as the capability of the underlying model.

Table 3: Contrasting designs for AI in learning, with likely effects.

Design feature	Answer-giving design	Learning-oriented design	Supporting evidence
Response to a request for help	Provides full solution	Gives hint or asks guiding question	Bastani et al. (2025)
Feedback granularity	Checks final answer	Responds to each reasoning step	VanLehn (2011)
Student engagement	Passive reading of answers	Constructive or interactive work	Chi & Wylie (2014)
Difficulty during practice	Reduced as far as possible	Kept at a productive level	Bjork (1994)
Teacher role	Bypassed	Designs prompts, monitors use	Bastani et al. (2025); Zawacki-Richter et al. (2019)

Note. Sources: [31], [16], [32], [33], [15].

Key takeaways.

- Generative AI can raise performance during practice without improving, and sometimes while harming, learning.
- AI tutors that guide rather than answer can avoid this harm.
- Long-standing principles of learning science should shape the design of AI tools.

CAN AI GIVE USEFUL FEEDBACK ON WRITING AND OPEN-ENDED WORK?

Feedback is among the most powerful influences on learning, but its effects vary widely. A review of feedback research concluded that feedback is most effective when it tells learners where they are going, how they are progressing and what to do next, and when it focuses on the task and the processes used to complete it rather than on the person [34]. A later meta-analysis confirmed that feedback has a substantial average effect on learning, with larger effects for information-rich feedback than for simple praise, rewards or punishment [35]. These findings imply that the value of AI feedback depends on its content, not merely on its speed.

Automated assessment of writing has a long history. In the 1960s Ellis Page predicted that computers would soon grade essays, and built a system that scored essays from surface features of the text [36]. Automated essay evaluation systems developed since then are used in large-scale testing and in writing instruction, and research has examined how closely their scores agree with those of human raters and how students use the feedback they provide [37]. A persistent criticism is that such systems reward features that are easy to measure, such as length and vocabulary, rather than the quality of argument, and that students can learn to write for the algorithm.

Large language models change the nature of automated feedback. They can comment on argument, organisation and evidence in natural language and respond to follow-up questions, which earlier systems could not do [6]. They also raise new concerns: their comments may be inaccurate or generic, they may rewrite students' work rather than help students improve it, and their judgements may differ across groups of writers in ways that reproduce bias [38]. The principles from feedback research offer a guide to design. AI feedback is most likely to help when it identifies specific next steps, asks students to revise their own work, and is combined with teacher judgement about what matters in a given task [34].

Key takeaways.

- Feedback helps most when it is specific, task-focused and points to next steps.
- Automated essay scoring has long been criticised for rewarding surface features.
- Language-model feedback is richer but must be designed to support revision, not replacement.

WHO BENEFITS, AND WHO MIGHT BE HARMED?

Equity

AI in education is often promoted as a way to reduce inequality by giving disadvantaged students access to personalised support. Some evidence supports this hope: adaptive software produced especially large gains for students who were furthest behind in Delhi [23]. But the benefits of technology depend on access to devices and connectivity, on students' self-regulation skills and on support from teachers and families, all of which are unequally distributed. Without deliberate design, technology can widen gaps by benefiting most those who are already advantaged.

Algorithmic bias

Predictive models in education can perform differently across groups. A review of algorithmic bias in education documented evidence of differences in predictive accuracy by race, ethnicity, gender, nationality and other characteristics across applications including dropout prediction, automated essay scoring and knowledge tracing [38]. Such disparities can arise from unrepresentative training data, from outcome measures that reflect existing inequalities, or from features that act as proxies for group membership. Researchers have argued that fairness in educational algorithms must be assessed against the specific decisions they inform, such as admissions, placement or support, and that different fairness criteria may conflict [39].

Privacy and student agency

Learning analytics depends on collecting detailed data about students, often without their active consent. Ethical analyses have highlighted questions about the ownership of data, the transparency of analytics to students, the risk that predictions become self-fulfilling labels, and the power imbalance between institutions and learners [40]. These authors argued that students should be treated as collaborators in analytics rather than merely as its subjects, with the ability to understand, and in some cases contest, how their data are used.

Table 4: Ethical issues in AI and learning analytics, with recommended safeguards.

Issue	Example concern	Recommended safeguard
Unequal access	Students without devices or connectivity cannot benefit	Provide access; evaluate effects by subgroup
Predictive bias	At-risk models less accurate for some groups	Audit accuracy and error rates by group
Labelling effects	Students treated differently once labelled at risk	Use predictions to offer support, not to restrict opportunity
Privacy	Extensive tracking of student activity	Data minimisation; clear policies; student consent where feasible
Over-reliance on AI	Students outsource thinking to chatbots	Learning-oriented AI design; teaching AI literacy
Teacher displacement	AI substitutes for rather than supports teachers	Involve teachers in design and deployment

Note. Sources: [38], [39], [40], [31], [15], [30].

Key takeaways.

- AI can narrow or widen educational gaps depending on access and design.
- Predictive models should be audited for differences in accuracy across groups.
- Students should understand and have a voice in how their data are used.

HOW DOES AI CHANGE THE WORK OF TEACHERS?

Discussions of AI in education often focus on students, but teachers determine whether and how technology is used in classrooms. Research on intelligent tutors increasingly treats teachers as partners rather than as obstacles. In one line of work, researchers designed tools that give teachers real-time information about students working with an intelligent tutor, such as which students are idle, struggling or misusing hints. In a classroom experiment, a mixed-reality tool that displayed this information to teachers through smart glasses led to greater learning gains, particularly for students who began with lower prior knowledge, because teachers directed their attention to the students who needed it most [41].

This research suggests a model of "teacher–AI complementarity", in which the AI handles tasks it does well, such as giving step-level feedback to many students at once, while teachers handle tasks that require judgement, motivation and relationships. The learning analytics cycle similarly depends on teachers to turn information into action [29]. Systematic reviews have nonetheless found that teachers are rarely involved in the design or evaluation of AI tools [15], and guidance from international organisations has emphasised the need to build teachers' capacity to use AI critically [30].

Generative AI also offers teachers direct support, for example in preparing materials, generating practice questions or drafting feedback [6]. These uses could free time for teaching, but they raise questions about accuracy, about the quality of AI-generated materials and about the deskilling of teaching if professional judgement is routinely delegated to machines. Research on these uses is at an early stage.

Key takeaways.

- Teachers are central to whether AI tools improve learning.
- Tools that inform teachers about students' progress can improve learning.
- Teachers are seldom involved in designing AI systems, which limits their fit with practice.

WHY IS THE EVIDENCE SO HARD TO INTERPRET?

Readers of the research on AI in education face several methodological problems that affect how results should be interpreted. The first concerns outcome measures. As the comparison between researcher-designed and standardised tests shows, effects measured on tests closely aligned with what a system teaches are often several times larger than effects on broad measures of achievement [19]. An analysis of hundreds of education studies found that effect sizes were systematically larger in small studies, in studies that used researcher-made measures and in quasi-experimental rather than randomised designs [42-62]. These features are common in evaluations of educational technology.

The second problem is duration and novelty. Many evaluations last only weeks, and short-term gains may reflect novelty or increased time on task rather than durable learning. The large-scale Cognitive Tutor trial showed that effects can change over the years of implementation [22], and the generative AI experiment showed that performance during practice can diverge sharply from later learning [31]. Evaluations should therefore measure learning after support is removed and over longer periods.

The third problem concerns the comparison condition. An AI tool compared with no instruction will appear more effective than the same tool compared with good teaching or with another well-designed intervention. Meta-analyses that separate comparison conditions, as Ma et al. did, give a more informative picture [18]. The fourth problem is publication and developer involvement: studies conducted by developers, and studies with positive results, may be more likely to be published. Independent evaluations, pre-registration and reporting of null results would strengthen the field [7, 42].

Key takeaways.

- Look for standardised outcomes, randomised designs and large samples.
- Measure learning after AI support is removed and over longer periods.

• Check what the comparison group received.

WHAT SHOULD TEACHERS, RESEARCHERS AND POLICYMAKERS DO?

For teachers and schools

The evidence suggests that teachers should treat AI tools as instruments whose value depends on how they are used. Tools that give step-level feedback and require students to do the thinking are more likely to support learning than tools that supply answers [16, 31]. Teachers can shape the use of AI by setting expectations, designing tasks that require students to explain their reasoning, and teaching students how to use AI as a learning aid rather than a shortcut. Implementation takes time, and early results may understate long-run effects [22].

For researchers

Research priorities include more randomised evaluations of AI tools in real classrooms, measurement of learning on outcomes that matter beyond the immediate task, and studies of long-term effects. Researchers should report results on standardised measures as well as on researcher-designed tests, since the difference can be large [7, 19]. They should also involve teachers in design and evaluation, as the systematic review of AI in higher education urged [15], and report effects separately for different groups of students.

For policymakers

Policymakers face pressure to adopt AI quickly. The evidence argues for a measured approach: investing in tools with demonstrated effects, requiring evidence of learning gains rather than of engagement or performance during use, protecting student data, and supporting teachers' professional development [8, 30]. Procurement processes can require vendors to provide evidence from rigorous evaluations and to make their systems auditable for bias.

Table 5: Summary of evidence by review question.

Question	Strength of evidence	Summary answer
Did the 2 sigma promise hold up?	Moderate	Human tutoring effects are large but well below two sigma; step-based ITS come close
How effective are ITS?	Strong	Meaningful gains in controlled studies; smaller on standardised tests and at scale
How do systems model knowledge?	Moderate	Knowledge tracing works; better prediction must translate into better teaching
Does learning analytics improve success?	Limited	Identifies at-risk students; causal evidence for interventions is scarce
What happens with generative AI?	Emerging	Unrestricted use can harm learning; guided designs can avoid harm

Question	Strength of evidence	Summary answer
Who benefits or is harmed?	Moderate	Effects vary by access and design; bias and privacy require safeguards

Note. Strength-of-evidence ratings are the authors' judgement based on the number and design of studies reviewed.

CONCLUSION

TheMachines can provide some of a teacher's most valuable functions, especially timely, specific feedback on students' reasoning, and, when they do, they can produce learning gains comparable to those of human tutors in some settings [16, 18]. Adaptive software has improved learning at scale in very different contexts [22, 23]. Learning analytics can help institutions identify students who need support [27].

But the evidence also shows that technology alone does not produce learning. Effects shrink on broad measures and at scale [19, 20]; predictions help only when they lead to effective action [29]; and the most powerful new tools can harm learning if they relieve students of the effort that learning requires [31]. The common thread is that pedagogical design, grounded in what is known about how people learn, determines whether AI helps or hinders [9, 32]. Bloom's question, how to give every student the benefit of a good tutor, remains open. AI offers new ways to pursue it, but only if those who build and use it keep the focus on learning rather than on what the technology can do.

REFERENECS

[1] Movva, C. K., & Manohar, V. (2021). OCF protocol implementation for smart home IoT applications: An Android SDK development case study. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 2(1), 89–96.

[2] Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education, 8*, 30–43.

[3] Potharaju, A. (2026). Human-guided artificial intelligence for adaptive compliance automation in financial institutions. *Journal of Computational Analysis and Applications*, 35(4), 57–68.

[4] Karim, N. B., Kovac, T., & Song-Chun, Z. (2025). Model Risk Management for Adaptive Credit Early-Warning Systems in Supervised SmallBusiness Lending: Aligning Machine-Learning Recalibration with Federal Model-Risk Guidance and the NIST AI Risk Management Framework. *International Journal of Acta Informatica*, *4*(1), 80-94.

[5] Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*(4), 253–278.

[6] Potharaju, B. (2026). Architecting national-scale regulatory reporting platforms: A high-availability infrastructure framework for compliance-critical systems. *Journal of International Crisis & Risk Communication*.

[7] Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review, 46*(5), 30–40.

[8] Mallempati, A. (2024). Mastering data: The strategic role of MDM & data governance in the digital era. *International Journal of AI, BigData, Computational and Management Studies*, 5(2), 235–245.

[9] Rahman, M. S., Bhowmik, P. K., Hossain, B., Tannier, N. R., Amjad, M. H. H., Chouksey, A., & Hossain, M. (2023). Enhancing fraud detection systems in the USA: A machine learning approach to identifying anomalous transactions. *Journal of Economics, Finance and Accounting Studies, 5*(5), 145-160.

[10] Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning analytics: From research to practice* (pp. 61–75). Springer.

[11] Potharaju, B., & Ravipudi, J. (2025). Automated data validation and resilient infrastructure design for enterprise regulatory reporting platforms. *International Journal of Computer Science and Engineering Innovations*, 1(1), 46–64.

[12] Karim, N. B., & Kovac, T. (2025). Cloud-Based Big-Data Architectures for AI-Driven Financial Planning and Analysis and Early Risk Signaling in Resource-Constrained Community Financial Institutions. *The Metascience, 3*(3), 29-40.

[13] Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274.

[14] Potharaju, A. (2026). Microservices-based modernization of financial regulatory systems. *International Journal of Computer Science and Engineering Innovations*, 2(1), 68–85.

[15] Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher, 49*(4), 241–253.

[16] Aluri, G. N. C., & Manohar, V. (2025). Generative AI integration patterns for enterprise workflow automation: A practitioner framework. *International Journal of Emerging Research in Engineering and Technology*, 6(3), 165–172.

[17] Escueta, M., Nickow, A. J., Oreopoulos, P., & Quan, V. (2020). Upgrading education with technology: Insights from experimental research. *Journal of Economic Literature, 58*(4), 897–996.

[18] Potharaju, A. (2026). Embedded compliance architecture for real-time financial systems modernization. *Journal of Information Systems Engineering and Management*, 11(2s), 1147–1158.

[19] Shammah, R. S., Sultana, S., Faruq, O., Kowser, S., Uddin, A., & Haque, S. A. (2026, July). Data-Driven Decision Support Framework for Engineering and IT Project Management Using Integrated Business Analytics and Management Information Systems. In *2026 IEEE International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)* (pp. 1-6). IEEE.

[20] Yuvaraj, N., Kumar, M. S., & Manoj, M. S. (2026). Unified agentic AI architecture for data lineage, governance, and predictive workflow automation in enterprise systems. In *2026 IEEE 18th International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 1817–1828). IEEE.

[21] Rashid, S. U., Al Muzakkiruddin, S., & Møller, C. (2026). Detecting and Correcting Temporal Drift in Credit-Risk and Fraud-Detection Models at Community Banks, Credit Unions, and CDFIs: An Automated Model-Risk-Monitoring Framework Below the Supervisory Threshold. *The Computertech*, 26-37.

[22] Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends, 59*(1), 64–71.

[23] Potharaju, B., & Sajja, R. R. (2024). Modernizing US federal regulatory reporting infrastructure for scalability, resilience, and audit readiness. *International Journal of Data Engineering and Intelligent Computing*, 7(1), 1–36.

[24] Aluri, G. N. C., & Mupparapu, H. K. (2021). Spring Batch framework patterns for large-scale data processing in cloud-native microservices: A GCP deployment study. *International Journal of AI, BigData, Computational and Management Studies*, 2(4), 130–142.

[25] Skinner, B. F. (1958). Teaching machines. *Science, 128*(3330), 969–977.

[26] Mallempati, A., & Jaladi, D. S. (2023). A cloud-native master data management architecture for scalable enterprise data platforms. *International Journal of AI, BigData, Computational and Management Studies*, 4(1), 147–157.

[27] Cuban, L. (2001). *Oversold and underused: Computers in the classroom*. Harvard University Press.

[28] Aluri, G. N. C., & Manohar, V. (2026). Model context protocol servers as orchestration layers for generative AI agents in enterprise software delivery. *International Journal of AI, BigData, Computational and Management Studies*, 7(2), 337–345.

[29] Reich, J., & Ruipérez-Valiente, J. A. (2019). The MOOC pivot. *Science, 363*(6423), 130–131.

[30] Potharaju, B., & Ravipudi, J. (2023). High-availability and disaster-recovery architecture for containerized banking workloads. *International Journal of Data Engineering and Intelligent Computing*, 6(1), 1–32.

[31] Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4–16.

[32] Karim, N. B., Tolosana, D., & Simko, R. (2025). Responsible and Legally Compliant Enterprise Artificial Intelligence in Small-Business Lending: Governance, Explainability, and Fair-Lending Safeguards for Early-Intervention Recommendation Engines. *International Journal of Modern Computing, 8*(1), 55-66.

[33] Yuvaraj, N., & Kumar, M. S. (2025). Agentic AI for zero-touch customer experience: A governance-constrained framework for autonomous service systems. *American International Journal of Computer Science and Technology*, 7(2), 100–111.

[34] Reich, J. (2020). *Failure to disrupt: Why technology alone can't transform education*. Harvard University Press.

[35] Kumar, M. S., Yuvaraj, N., & Manoj, M. S. (2026). Intelligent orchestration of master data and privacy compliance via agentic AI for customer lifecycle optimization. In *2026 IEEE 18th International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 1715–1723). IEEE.

[36] Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.

[37] Kumar, M. S., & Yuvaraj, N. (2024). Predictive customer experience orchestration using governed data pipelines and intelligent service signals. *International Journal of Emerging Trends in Computer Science and Information Technology*, 5(1), 206–215.

[38] Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education, 16*, 39.

[39] Mallempati, A., Jaladi, D. S., & Jegajothi, B. (2026). HDLF-DGC: A hybrid deep learning framework for intelligent cybersecurity and data governance in distributed web systems. In *2026 Third International Conference on Innovations in Cybersecurity and Data Science (ICICDS)* (pp. 16–21). IEEE.

[40] VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221.

[41] Jaladi, D. S., Mallempati, A., & Jegajothi, B. (2026). A hybrid federated privacy-aware deep learning framework for secure data governance in cloud systems. In *2026 4th International Conference on Sustainable Computing and Smart Systems (ICSCSS)* (pp. 1629–1636). IEEE.

[42] 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.

[43] Potharaju, A. (2026). Policy-driven authorization architectures for modern financial regulatory platforms. *Journal of International Crisis and Risk Communication Research*, 9(4), 36–44.

[44] Rahman, M. S., Kabir, A. A., Reon, M. A. H., & Rashid, S. U. (2025). From Financial Transactions to Clinical Records: A Unified Machine Learning Methodology for Real-Time Anomaly Detection and Institutional Decision Support In High-Stakes Domains. *International Journal of Information Technology and Management Information Systems (IJITMIS), 16*(6), 23-42.

[45] Potharaju, A. (2021). Designing dynamic questionnaire engines for risk-based account review workflows. *International Journal of Applied Data Science & Modern Computing*, 4(1), 1–37.

[46] Siddiqui, M. I. H., Rahman, M. S., Kabir, A. A., Mahmud, F. U., Rashid, S. U., & Shammah, R. S. (2023). Comparative analysis of explainable machine learning models for cancer classification using cytological features. *Journal of Medical and Health Studies, 4*(5), 110-150.

[47] 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.

[48] Movva, C. K., & Aluri, G. N. C. (2025). Enterprise mobile payment gateway architecture on IoT handheld devices: Security and reliability design. *International Journal of AI, BigData, Computational and Management Studies*, 6(1), 232–241.

[49] Shammah, R. S., Rahman, I., Guria, Z. M., Kowser, S., Faruq, O., & Morshed, N. (2025, December). Scalable Machine Learning Architecture for Securing IoT Networks with Digital Governance in the AI Era of Cyber-Resilient Smart Cities. In *2025 International Conference on AI-Driven STEM Education and Learning Technologies (AISTEMEDU)* (pp. 1-7). IEEE.

[50] Potharaju, B. (2026). Infrastructure-level intelligence: Embedding AI into data movement and validation layers. *International Journal of Computational and Experimental Science and Engineering*.

[51] Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research, 86*(1), 42–78.

[52] Shahiduzzaman, M., Sourav, M. S. A., Kabtia, M., Shammah, R. S., Islam, M. A., Chakraborty, P., ... & Adnan, M. (2026). Attention-guided quantum-inspired extreme learning machine with grey wolf optimization for rare-attack detection in IIoT networks. *Discover Artificial Intelligence*.

[53] Steenbergen-Hu, S., & Cooper, H. (2013). A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students' mathematical learning. *Journal of Educational Psychology, 105*(4), 970–987.

[54] Movva, C. K., & Mupparapu, H. K. (2026). MVVM architectural patterns for enterprise Android kiosk applications: State management and lifecycle handling. *International Journal of Emerging Trends in Computer Science and Information Technology*, 7(2), 320–330.

[55] Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology, 106*(2), 331–347.

[56] Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014). Effectiveness of Cognitive Tutor Algebra I at scale. *Educational Evaluation and Policy Analysis, 36*(2), 127–144.

[57] Zuberi, N., Rahman, M. M., Rashid, S. U., Ahsan, M. N., Rahman, M. D., & Mollah, M. H. O. R. (2026). A robust machine learning framework for detecting temporal drift in financial fraud prevention. *Scientific Reports*.

[58] Mupparapu, H. K., & Movva, C. K. (2020). Serverless migration patterns for compliance-critical financial applications on Microsoft Azure. *International Journal of Emerging Research in Engineering and Technology*, 1(2), 85–96.

[59] Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting education? Experimental evidence on technology-aided instruction in India. *American Economic Review, 109*(4), 1426–1460.

[60] Karim, N. B., & Moreau, L. (2026). Quantifying the Prospective Impact of AI-Assisted Early Intervention in US Small-Business Lending: Default Avoidance, Credit-Access Preservation, and Small-Business Employment Effects Across Community Lending Portfolios. *The Metascience*, *4*(2), 33-44.

[61] Rahman, M. M., Computer, M. M. R., Das, N., Uddin, M. S., Shammah, R. S., & Rahman, M. M. (2025, May). An Investigation of Face Detection and Feature Extraction Techniques in Deepfake Detection Using a Multi-Model Machine Learning Approach. In *2025 3rd International Conference on Business Analytics for Technology and Security (ICBATS)* (pp. 1-8). IEEE.

[62] Manohar, V., & Mupparapu, H. K. (2025). Scalable data pipeline architecture for real-time supply chain analytics using PySpark and Snowflake. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 6(3), 203–212.