# THE PROMISE AND THE EVIDENCE

*Educational Technology and AI in Learning*

Applied Pedagogy Research Lab

*Guido Bartolucci, Principal Investigator*
guido@appliedpedagogy.com
LAB.APPLIEDPEDAGOGY.COM

L1-005 · March 2026

*Research conducted by AI agents (Claude, Anthropic) under human direction.*
*See* LAB.APPLIEDPEDAGOGY.COM *for methodology and verification framework.*

# CONTENTS

I FOUNDATIONS

II THE AI FRONTIER

III FAILURE MODES AND SYNTHESIS

# Part I

# FOUNDATIONS

# 1

# THE CENTRAL TENSION

Educational technology has been promising to transform learning for the better part of a century. Radio would bring the world's best teachers to every classroom. Television would do the same but better. Computers would personalize instruction. The internet would democratize knowledge. MOOCs would make elite education free. Virtual reality would make learning immersive. And now, large language models will revolutionize tutoring.

Each wave follows the same arc: extravagant promises, enthusiastic adoption, disappointing evaluations, and eventual integration as one tool among many. This pattern — what Gartner calls the hype cycle and what education researchers know from bitter experience — is not evidence that technology is useless. It is evidence that technology is a delivery mechanism, not a pedagogy. The same principles of effective instruction apply regardless of whether the medium is a chalkboard, a screen, or a conversational AI: manage cognitive load, provide retrieval practice, give timely and specific feedback, scaffold appropriately for the learner's current expertise, and support autonomy rather than compliance.

This investigation examines what educational technology actually delivers, evaluated against both the empirical record and the five-layer competence stack that Applied Pedagogy has adopted as its outcome specification. The competence stack — domain knowledge, skill, judgment, metacognition, and character — provides the evaluative framework throughout. Technologies that address only layers 1–2 while ignoring layers 3–5 are insufficient by definition, regardless of how effectively they deliver content.

The investigation proceeds in five parts. First, the intelligent tutoring systems (ITS) literature, which represents decades of rigorous research on automated instruction. Second, Mayer's multimedia learning principles, which provide the strongest theoretical framework for digital learning design. Third, the emerging literature on AI and large language models in education — where the evidence is thinnest and the stakes are highest. Fourth, the failure modes that technology enthusiasts rarely discuss. Fifth, a synthesis that maps technologies onto the competence stack and derives design principles for Applied Pedagogy's use of technology.

The defining challenge is epistemic honesty in the face of hype. The AI-in-education space is dominated by vendor claims and small pilot studies. The job is to find whatever rigorous evidence exists, apply the principles established by decades of cognitive science research, and be transparent about the boundary between evidence and extrapolation.

# 2 WHAT DECADES OF ITS RESEARCH ACTUALLY SHOW

## 2.1 THE TWO-SIGMA PROBLEM AND ITS DISCONTENTS

In 1984, Benjamin Bloom reported that students receiving one-on-one tutoring from expert human tutors performed two standard deviations above conventionally taught students — the famous “two-sigma problem.” If true, this would mean the average tutored student outperformed 98% of conventionally taught students, an effect size so large it would dwarf nearly every other educational intervention. Bloom framed this as a challenge: could any scalable intervention match what individual tutoring achieves?

The two-sigma finding has driven thirty years of ITS research. But Bloom’s original study has been questioned. The two-sigma effect was observed under ideal conditions — expert tutors working one-on-one with highly motivated students — and subsequent research has found that typical human tutoring produces more modest effects. VanLehn (2011), in the most important comparative review of its kind, found that human tutoring produced an average effect size of $d \approx 0.79$ compared to no tutoring. This is large by educational standards but considerably less than two sigma. The gap between Bloom’s claim and VanLehn’s finding matters because it recalibrates expectations for what ITS need to achieve.

## 2.2 THE VANLEHN REVIEW: STEP-LEVEL VERSUS PROBLEM-LEVEL

VanLehn’s (2011) analysis introduced a distinction that transforms the conversation about ITS effectiveness. He separated tutoring systems by the grain size at which they interact with learners. “Problem-level” systems evaluate student work only after a complete problem is submitted — they provide help and feedback at the level of whole solutions. “Step-level” systems interact with students during problem-solving, evaluating each step and providing feedback within the solution process.

The finding is striking. Problem-level tutoring systems — including many popular commercial systems — produced effect sizes statistically indistinguishable from zero when compared to human tutoring. They added little beyond what a well-designed workbook with an answer key could provide. Step-level tutoring systems, by contrast, produced effect sizes comparable to human tutoring ($d \approx 0.76$). The mechanism is clear: effective tutoring — whether human or automated — requires engagement with the learner’s reasoning process, not just their final answers. This maps directly to what L1-003 found about feedback design: task-level and process-level feedback improve learning, while outcome-only feedback does not.

The implication is that the “inner loop” of tutoring — the moment-by-moment interaction between the learner’s thinking and the tutor’s response — is where the learning happens. Systems that wait for a complete answer before responding miss the opportunity to catch and correct misconceptions in real time, to redirect unproductive strategies, and to scaffold the specific step where the learner is stuck.

## 2.3 META-ANALYTIC EVIDENCE

Two major meta-analyses quantify ITS effectiveness. Ma, Adesope, Nesbit, and Liu (2014) conducted a meta-analysis published in the *Journal of Educational Psychology* that examined ITS effects across multiple outcome measures. They found that ITS produced moderate positive effects on learning outcomes, with effect sizes varying by the type of comparison condition and the domain. When compared to conventional classroom instruction, ITS showed a meaningful advantage; when compared to individual human tutoring, the advantage was small or absent.

Kulik and Fletcher (2016) conducted a separate meta-analytic review focusing specifically on ITS effectiveness, finding an overall effect size of approximately d ≈ 0.66 compared to conventional instruction. This is a meaningful effect — roughly equivalent to moving a student from the 50th to the 75th percentile — but it falls well short of Bloom's two-sigma promise. The moderator analysis is instructive: effects were larger for well-structured domains (mathematics, physics, programming) and smaller for ill-structured domains (writing, social studies). This suggests that ITS effectiveness is partly a function of domain structure, not just system quality.

## 2.4 WHAT MAKES ITS WORK

When ITS are effective, what features drive the effect? The research points to several mechanisms, all of which reflect established principles from cognitive science and instructional design:

**Adaptive problem selection.** Effective ITS choose problems at the edge of the learner's current competence — challenging enough to produce learning, not so difficult as to produce frustration. This is mastery learning implemented computationally: the system monitors student performance and advances to new material only when current material is mastered. The connection to cognitive load theory (L1-004) is direct: adaptive problem selection keeps intrinsic load at manageable levels while minimizing extraneous load from problems that are too easy or too hard.

**Inner-loop feedback.** As VanLehn demonstrated, the critical feature is step-level engagement. The most effective ITS — Cognitive Tutor, ANDES, AutoTutor — provide feedback on individual solution steps, catching errors and misconceptions as they emerge. AutoTutor, developed by Graesser and colleagues, uses natural language dialogue to engage students in explanatory reasoning, asking questions like "What would happen if…" and "Can you explain why…" (Nye, Graesser & Hu, 2014). This connects to L1-003's finding that feedback should be process-focused and task-focused, not person-focused.

**Mastery learning.** Many effective ITS implement mastery-based progression: students work on a topic until they demonstrate competence, then move on. This prevents the accumulation of knowledge gaps that plagues time-based progression systems. Koedinger, Corbett, and Perfetti (2012) formalized this in their Knowledge-Learning-Instruction framework, which maps between knowledge types, learning processes, and instructional events to optimize learning efficiency.

**Bug analysis and misconception targeting.** The most sophisticated ITS maintain models of common student misconceptions ("bugs") and design feedback to address specific misunderstandings rather than providing generic correction. This requires extensive cognitive task analysis of the domain — mapping the typical errors students make and the conceptual misunderstandings that produce them.

## 2.5 THE LIMITATIONS OF ITS

Despite decades of development, ITS face fundamental limitations that the research clearly documents:

**Domain restriction.** ITS work best in well-structured domains where problems have definable solution paths and correctness can be algorithmically assessed. Mathematics, physics, programming, circuit design, medical diagnosis — these domains have clear right answers and specifiable solution procedures. The ITS literature has far less to show for ill-structured domains like writing, ethical reasoning, historical interpretation, or creative problem-solving. Baker (2016), in a provocatively titled paper "Stupid Tutoring Systems, Intelligent Humans," argued that ITS should focus on what they do well — structured practice with feedback — and leave the more complex aspects of learning to human teachers.

**Layer limitation.** Evaluated against the competence stack, ITS primarily address layers 1 and 2: domain knowledge acquisition and skill development. They can drill facts, practice procedures, and provide corrective feedback on well-defined tasks. But they struggle with layer 3 (judgment) because judgment requires exposure to ambiguous, ill-structured situations where multiple reasonable approaches exist. They rarely address layer 4 (metacognition) beyond basic confidence calibration. And they cannot address layer 5 (character and disposition) at all — the epistemic virtues of intellectual honesty, tolerance for uncertainty, and willingness to say "I don't know" require a social and environmental context that software cannot provide.

**The floor under human tutoring.** Even the best ITS have not surpassed human tutoring for complex learning outcomes. VanLehn's finding that step-level ITS match human tutoring applies to structured domains where the learning goal is skill acquisition. For developing judgment, supporting identity formation, or building relationships that sustain long-term motivation, human tutors remain irreplaceable.

**Cost and development time.** Building effective ITS requires extensive cognitive task analysis, student modeling, and iterative testing. The development cost per hour of instruction is orders of magnitude higher than conventional materials. This has limited ITS deployment to high-value, high-volume domains where the investment can be amortized across millions of users.

## 2.6 THE ITS-TO-LLM TRANSITION

The history of ITS is directly relevant to the current LLM moment because LLMs possess several capabilities that traditional ITS lacked — and these capabilities fundamentally change the design space.

Traditional ITS required years of domain modeling to function: cognitive task analysis of the domain, explicit encoding of expert solution paths, cataloging of common student "bugs" and misconceptions, and design of feedback strategies for each. This made ITS development prohibitively expensive for most educational contexts and confined effective systems to a handful of well-studied domains — primarily mathematics, physics, and programming. The Cognitive Tutor, for example, required over a decade of development drawing on ACT-R cognitive architecture research before achieving its documented effectiveness.

LLMs bypass this bottleneck entirely. They can generate explanations, evaluate solutions, and provide feedback across virtually any domain without requiring explicit domain modeling. This makes them vastly more scalable but potentially less precise — a traditional ITS that detected a specific misconception could provide targeted corrective feedback designed by domain experts, while an LLM generates feedback based on statistical patterns in its training data.

The question is whether the flexibility and scalability of LLMs compensate for their lack of precise student modeling. Roll and Wylie (2016) argued that the evolution of AIED should move toward "revolution" — fundamentally rethinking what AI can contribute to education rather than incrementally improving existing paradigms. LLMs may represent exactly this kind of revolution, but the ITS literature provides a crucial caution: the features that made ITS effective (step-level engagement, process-focused feedback, mastery learning, bug analysis) are not incidental — they reflect deep principles of how learning works. Any LLM-based educational tool that ignores these principles in favor of superficially impressive conversational capability risks repeating the mistake of problem-level ITS: producing engaging but pedagogically shallow interactions.

Stamper, Xiao, and Hou (2024) made this argument explicitly, contending that decades of ITS research should inform the design of LLM-based educational tools. The danger is that the AI community, dazzled by the conversational capabilities of LLMs, will ignore the hard-won lessons of the ITS community about what makes automated instruction actually effective. Holstein, McLaren, and Aleven (2019) offered a complementary perspective: rather than AI replacing teachers, the most productive approach may be AI supporting teachers — providing real-time information about student learning that enables more effective human instruction. This "teacher-AI complementarity" model positions technology as a tool for human teaching rather than a substitute for it.

## 2.7 ESCUETA'S ECONOMIC LENS

Escueta, Nickow, Oreopoulos, and Quan (2020), writing in the *Journal of Economic Literature*, brought an economist's rigor to the educational technology evidence base. Their review focused exclusively on experimental and quasi-experimental evidence — a much higher methodological bar than most ed-tech reviews. Their key finding: the technology interventions with the strongest effects were those that implemented proven pedagogical principles (spaced practice, immediate feedback, personalized pacing) rather than those that introduced novel technologies for their own sake. Technology that made existing good practices more efficient and scalable produced learning gains; technology that was deployed without pedagogical grounding did not.

This finding reinforces the delivery-mechanism framing: technology is a tool for implementing pedagogy, not a replacement for it. The most effective educational technologies are, in a sense, invisible — they implement spacing, retrieval, feedback, and adaptation so seamlessly that the learner experiences effective instruction without being aware of the technology's role. The least effective are the opposite — technologically impressive, visually engaging, and pedagogically empty.

## 2.8 THE TWO-SIGMA PROBLEM REVISITED

Bloom's challenge was to find scalable interventions that match the effectiveness of individual tutoring. After thirty years, the honest answer is: ITS get roughly halfway there. With d ≈ 0.66–0.76 in structured domains, they close a significant portion of the gap between conventional instruction and human tutoring. But the remaining gap is exactly in the areas that technology handles poorly: the responsive, relationship-based, adaptive qualities of human interaction that address the upper layers of the competence stack.

The ITS literature thus provides a clear and sober baseline for evaluating AI-based educational tools. Any new technology — including LLM-based tutors — should be compared not to conventional instruction (a low bar) but to the best ITS and to human tutoring. And the comparison should include not just knowledge acquisition (layer 1–2) but judgment, metacognition, and the dispositional outcomes that constitute full competence.

# 3 MULTIMEDIA LEARNING PRINCIPLES: THE DESIGN CONSTRAINTS

## 3.1 MAYER'S FRAMEWORK

Richard Mayer's cognitive theory of multimedia learning, first systematized in 2002 and updated through a third edition in 2020, provides the strongest theoretical framework for educational technology design. The theory rests on three cognitive science assumptions: (1) humans process visual and auditory information through separate channels (dual coding), (2) each channel has limited capacity (cognitive load), and (3) meaningful learning requires active cognitive processing — selecting, organizing, and integrating information.

From these assumptions, Mayer derived a set of principles, each supported by multiple experiments, that specify how to design instructional multimedia to maximize learning. The principles are not merely theoretical — they are engineering constraints. Any educational technology that violates them will, according to the theory and the evidence, produce less learning than one that observes them. The key principles include:

**The coherence principle.** People learn better when extraneous material is excluded rather than included. Decorative graphics, background music, irrelevant animations, and "interesting but irrelevant" details all reduce learning by consuming limited cognitive resources with material that does not support the instructional objective. This principle is the most commonly violated in commercial ed-tech, which routinely adds visual complexity for engagement purposes that actively impedes learning.

**The signaling principle.** People learn better when cues are added that highlight the organization of essential material — headings, arrows, highlighting of key terms, verbal emphasis. Signaling reduces extraneous processing by directing attention to what matters.

**The redundancy principle.** People learn better from graphics and narration than from graphics, narration, and on-screen text. Adding redundant text to a narrated animation forces learners to process the same information through both channels simultaneously, creating extraneous load rather than reducing it. This principle is routinely violated by educational videos that display text while the narrator reads it aloud.

**The spatial contiguity principle.** People learn better when corresponding words and pictures are presented near each other rather than far apart on the page or screen. When a diagram is on one page and the explanation on another, learners waste cognitive resources searching for the relevant connections.

**The temporal contiguity principle.** People learn better when corresponding narration and animation are presented simultaneously rather than successively. Presenting all the animation first and then the narration (or vice versa) forces learners to hold information in working memory while waiting for the corresponding material.

**The segmenting principle.** People learn better when a complex lesson is presented in learner-paced segments rather than as a continuous unit. This allows learners to fully process each segment before moving to the next, preventing cognitive overload.

**The pre-training principle.** People learn better from a multimedia lesson when they first receive pre-training that introduces the names and characteristics of key concepts. Pre-training reduces intrinsic load during the main lesson by ensuring that learners already have the basic schema needed to process the material.

**The modality principle.** People learn better from graphics and narration than from graphics and on-screen text. Narration uses the auditory channel, leaving the visual channel free for processing graphics. On-screen text competes with graphics for visual processing capacity.

**The personalization principle.** People learn better when words are in conversational style rather than formal style. Using "you" and "I" rather than third-person formal language promotes a sense of social partnership that increases cognitive engagement.

## 3.2 WHEN MORE TECHNOLOGY MEANS LESS LEARNING

The most important finding for the current AI moment comes from Makransky and colleagues' research on immersive virtual reality in education. Makransky, Terkildsen, and Mayer (2019) conducted a study comparing a science lab simulation delivered via immersive VR headset versus desktop computer. The VR condition produced significantly more presence (the feeling of "being there") and higher self-reported engagement — but significantly less learning. The explanation is straightforward in terms of cognitive load theory: the VR technology itself consumed cognitive resources that were then unavailable for processing the instructional content. The sensory richness of the immersive environment was extraneous load masquerading as engagement.

Makransky and Petersen (2021) formalized this insight in their Cognitive Affective Model of Immersive Learning (CAMIL), which distinguishes between the affective and cognitive pathways through which immersive technology influences learning. Presence and embodiment can increase motivation and emotional engagement (the affective pathway), but they can also increase extraneous cognitive load and reduce learning (the cognitive pathway). The net effect depends on whether the immersive features are integral to the learning objective — learning to perform surgery might genuinely benefit from spatial immersion, while learning the periodic table does not.

The lesson generalizes far beyond VR. Any technology that increases engagement through sensory novelty, gamification, or interactivity runs the risk of optimizing for time-on-task and subjective engagement while reducing actual learning. The students feel like they are learning more; they are learning less. This is the engagement trap, and it is one of the most consequential risks of the current technology-in-education moment.

Makransky and Mayer (2022) subsequently proposed an "immersion principle" — that immersive VR can enhance learning outcomes, but only when two conditions are met: (1) the immersion is integral to the learning objective (e.g., spatial navigation, physical procedure), and (2) generative learning strategies (self-explanation, summarization, teaching) are incorporated to promote active processing. Without these strategies, the technology absorbs cognitive resources without producing corresponding learning gains.

## 3.3 THE EVIDENCE BASE FOR MAYER'S PRINCIPLES

The strength of Mayer's framework lies in its extensive empirical foundation. Unlike many educational theories that rest on a handful of studies, Mayer's principles have been tested across hundreds of experiments, multiple research groups, diverse populations, and varied domains. The coherence principle alone has been supported by over a dozen controlled experiments showing that adding interesting but irrelevant material to a multimedia lesson consistently reduces learning — even when (especially when) the added material is engaging and students report enjoying it.

The redundancy principle has been replicated enough to qualify as one of the more robust findings in educational technology research, though with important boundary conditions. Kalyuga and Sweller (2014) demonstrated that redundancy effects depend on learner expertise: for novices processing unfamiliar material, simultaneous text and narration is harmful; for more advanced learners who already have relevant schemas, redundant information may be processed selectively without the same cost. This expertise-dependent effect parallels the expertise reversal effect documented in L1-004's instructional design literature — what helps novices can hurt experts, and vice versa.

The pre-training principle connects directly to Kapur's productive failure research. Both recognize that learners need activation of relevant prior knowledge before they can meaningfully process complex new information. Where they differ is in the mechanism: Mayer's pre-training provides the prior knowledge directly through advance instruction, while Kapur's productive failure activates whatever prior knowledge the learner already has through struggle. These are complementary approaches — pre-training may be more appropriate when learners lack the basic vocabulary to engage with a problem, while productive failure may be more appropriate when learners have some foundation but need to discover its limits.

Meyer, Omdahl, and Makransky (2019) demonstrated that the pre-training principle applies to immersive VR learning — students who received pre-training before a VR experience showed better learning outcomes than those who did not, because pre-training reduced the cognitive load during the immersive experience. This has direct implications for any technology that creates high cognitive demand through its interface: if the medium is demanding, pre-training becomes even more important.

## 3.4 THE FLIPPED CLASSROOM AS MULTIMEDIA APPLICATION

The flipped classroom model — where students watch video lectures at home and use class time for active learning — can be understood as an application of several multimedia principles simultaneously. Students watch segmented, learner-paced videos (segmenting principle), can pause and rewind as needed (reducing cognitive overload), and then apply what they've learned through active practice with teacher support (which leverages the testing effect and feedback).

The evidence on flipped classrooms is moderately positive. Hew and Lo (2018) conducted a meta-analysis focused on health professions education and found flipped classrooms produced small to moderate improvements in learning outcomes compared to traditional lectures. But the evidence is messy because "flipped classroom" implementations vary enormously — some are well-designed multimedia presentations followed by structured active learning, while others are hastily recorded lectures followed by unstructured class time. The pedagogical design matters far more than the label.

This connects to a broader point about educational technology research: the effect of a technology depends almost entirely on how it is used, not on what it is. A poorly designed flipped classroom can be worse than a good lecture; a well-designed lecture can be better than a poorly designed simulation. The technology is a delivery mechanism; the pedagogy is the active ingredient.

## 3.5 APPLYING MAYER TO MODERN ED-TECH

Most commercial educational technology violates Mayer's principles routinely:

- **Learning apps** add decorative animations, sound effects, and gamified reward sequences that violate coherence.

- **Educational videos** display text while narrating it aloud, violating redundancy.
- **Online courses** present hour-long lectures without segmenting or learner pacing.
- **Interactive simulations** add complexity for engagement without ensuring it serves learning objectives.

The practical implication is that instructional designers should audit every digital learning tool against Mayer's principles before deployment. A beautiful, engaging, technologically impressive tool that violates coherence, redundancy, and segmenting will produce less learning than a plain text document with a few well-designed diagrams.

# 4 SPACED REPETITION SOFTWARE: TECHNOLOGY THAT WORKS

## 4.1 THE COGNITIVE SCIENCE FOUNDATION

If there is one area where technology straightforwardly implements a well-established cognitive science principle, it is spaced repetition. The spacing effect — that distributed practice produces better long-term retention than massed practice — has been replicated continuously since Ebbinghaus first demonstrated it in 1885. It is one of the most robust findings in all of psychology. The testing effect — that retrieval practice strengthens memory more than re-reading — is similarly well-established (L1-003 found it to be "one of the most robust findings in cognitive psychology" with d ≈ 0.5 in classroom settings).

Spaced repetition software (SRS) — Anki, SuperMemo, Duolingo's internal algorithm, and similar systems — combines these two effects computationally. The software schedules review of each item at increasing intervals, calibrated to the point where the learner is about to forget. Each successful retrieval strengthens the memory and extends the interval; each failure shortens it. The result is a personalized spacing schedule that keeps each piece of knowledge at the threshold of retrieval — maximally effortful (which maximizes learning) without being so effortful that retrieval fails.

## 4.2 THE EVIDENCE

Settles and Meeder (2016) described Duolingo's half-life regression model for spaced repetition in language learning, demonstrating that computationally optimized spacing algorithms can adapt to individual forgetting rates. Kornell (2009) showed that even simple flashcard-based spaced retrieval produces significant learning gains, and that spacing across multiple sessions outperforms cramming within a single session — a finding that holds even though students consistently *believe* massed practice is more effective (the metacognitive error that L1-002 identified).

More recent evidence comes from medical education, where Anki use has become widespread. Gilbert et al. (2023) conducted a cohort study examining the impact of Anki on medical school academic performance and found positive associations between SRS use and exam scores, though the observational design limits causal inference. Jape, Zhou, and Bullock (2022) found that a spaced repetition intervention enhanced both learning outcomes and student engagement in medical pharmacology.

## 4.3 THE METACOGNITIVE ERROR AND SRS DESIGN

There is a subtlety in the SRS evidence that connects directly to L1-002's findings on metacognitive errors. Kornell (2009) found that students who used massed practice (cramming) believed they had learned more than students who used spaced practice, despite the spaced practice group demonstrating superior retention on delayed tests. The subjective experience of learning — the feeling of fluency that comes from just having reviewed material — is misleading. Massed practice creates a strong feeling of knowing; spaced practice creates a weaker feeling but stronger actual retention.

SRS software addresses this metacognitive error structurally. By scheduling reviews at intervals calibrated to the forgetting curve, the software imposes spaced practice regardless of the learner's preferences. The learner does not need to overcome the natural inclination toward massed practice because the algorithm manages the spacing. This is an example of technology implementing a learning science principle *better than the learner would choose to implement it themselves* — precisely because humans are systematically miscalibrated about what learning strategies are most effective.

The implications extend beyond SRS. In many contexts, the learner's subjective experience of "what works" is a poor guide to actual effectiveness. Students prefer highlighting over retrieval practice, massed over spaced review, being told the answer over struggling to recall it — and in every case, the preferred strategy produces less learning. Technology that gives learners complete control over their learning strategy may inadvertently optimize for the subjective experience of learning rather than learning itself. Conversely, technology that imposes effective strategies — even over learner objections — may produce better outcomes at the cost of perceived autonomy.

This creates a tension with L1-002's emphasis on autonomy support. SDT holds that autonomy is a basic psychological need whose frustration undermines intrinsic motivation. But if learners autonomously choose ineffective strategies, honoring their autonomy comes at the cost of their learning. The resolution, as L1-004 suggested, is providing autonomy within structure: learners can choose *what* to study and *when*, but the *how* (spacing, retrieval practice, interleaving) is built into the system's architecture. Choice within constraints, not unlimited choice.

## 4.4 STRENGTHS AND LIMITATIONS

Spaced repetition software genuinely works for its target use case: layer 1 knowledge acquisition — vocabulary, facts, definitions, procedural steps, diagnostic criteria. It implements a proven cognitive principle more efficiently than any human teacher could, personalizing review schedules across thousands of items for each individual learner.

But SRS addresses only a narrow band of the competence stack. It is optimized for declarative knowledge retrieval — knowing *that*. It cannot develop skill (layer 2) because skill requires contextualized practice, not decontextualized recall. It cannot develop judgment (layer 3) because judgment requires exposure to varied, ambiguous situations, not isolated fact retrieval. It cannot develop metacognition (layer 4) beyond the minimal self-monitoring involved in rating one's confidence during retrieval. And it has no mechanism for addressing character (layer 5).

The risk is that SRS becomes the model for what educational technology should be — that the success of spaced repetition for factual knowledge is taken as evidence that technology can address the full range of learning objectives. It cannot. SRS is a precision tool for one specific layer of one specific type of learning. Its success should not be generalized.

Part II

# THE AI FRONTIER

# 5 AI AND LARGE LANGUAGE MODELS IN EDUCATION: WHAT WE KNOW, WHAT WE DON'T, AND WHAT COGNITIVE SCIENCE PREDICTS

## 5.1 THE CURRENT EVIDENCE BASE

This section requires an epistemic health warning at the outset. As of early 2025 — less than two and a half years after ChatGPT's public release in November 2022 — the evidence base for LLM-based educational tools consists almost entirely of opinion pieces, conceptual frameworks, small pilot studies, and surveys of student and teacher perceptions. Rigorous, controlled outcome studies with adequate sample sizes and validated measures are nearly nonexistent. The literature is dominated by what Zawacki-Richter et al. (2019) had already identified in the pre-LLM AIED literature: a field where computer scientists and AI researchers publish enthusiastic work that educators are largely absent from.

Kasneci et al. (2023) published the most widely cited early review, with over 4,300 citations in under two years — a citation velocity that reflects the intensity of interest but not the maturity of the evidence. Their review identified opportunities (personalized learning, intelligent tutoring, assessment support, administrative efficiency) and challenges (bias, privacy, academic integrity, over-reliance) — a reasonable mapping of the landscape but one based primarily on theoretical analysis rather than outcome data.

Lo (2023) conducted a rapid review of the ChatGPT-in-education literature and found it dominated by opinion pieces and early-stage empirical work. The few studies with learning outcome measures were small-scale and short-duration. Farrokhnia et al. (2023) performed a SWOT analysis reaching similar conclusions: substantial theoretical promise, minimal empirical validation.

Yan et al. (2023) conducted a scoping review of practical and ethical challenges of LLMs in education, identifying concerns about accuracy, bias, privacy, academic integrity, and the potential for undermining critical thinking. Gordon et al. (2024), in a systematic BEME Guide reviewing AI in medical education, found that the majority of studies were descriptive or single-institution pilots, with few randomized controlled trials and limited evidence of impact on actual patient care or clinical reasoning.

The most methodologically rigorous study I encountered is Darvishi et al. (2023), published in *Computers & Education* with an FWCI of 88.50. This study examined the impact of AI assistance on student agency and found that while AI tools could support learning in some contexts, they also risked reducing student agency — the sense of ownership and control over one's learning process. This connects directly to L1-002's finding that autonomy support is essential for intrinsic motivation: tools that do the thinking for students may produce correct answers while undermining the motivational and metacognitive processes that produce genuine learning.

## 5.2 THE EDUCATOR-ABSENCE PROBLEM

Zawacki-Richter et al.'s (2019) finding that educators are largely absent from AIED research deserves extended discussion because it illuminates a structural problem that persists into the LLM era. Their systematic review of 146 studies on AI applications in higher education found that the research was dominated by computer science and engineering perspectives. Learning scientists, curriculum designers, educational psychologists, and practicing teachers were rarely authors, rarely consulted, and rarely cited. The result was technically sophisticated tools designed without reference to the principles that govern how humans actually learn.

This pattern has, if anything, intensified with the LLM revolution. The development of ChatGPT, Claude, and similar models was driven by AI researchers focused on capability benchmarks — reasoning, knowledge recall, linguistic fluency — not on pedagogical effectiveness. The result is tools that are impressively capable as general-purpose AI systems but not designed with learning science principles in mind. They default to answer-giving mode because that is what users request and what capability benchmarks reward. But answer-giving is precisely the interaction pattern that the ITS literature has shown to be least effective for learning.

The gap between AI capability and pedagogical design is not merely an academic concern. It means that the LLM-based educational tools now being deployed in classrooms, universities, and corporate training were not designed by people who understand the testing effect, productive failure, autonomy support, or the distinction between step-level and problem-level feedback. They were designed by people who understand transformer architectures, reinforcement learning from human feedback, and prompt engineering. These are different skill sets, and the absence of the former in the design process is a liability that the field has not yet seriously addressed.

## 5.3 THE SCALE OF DEPLOYMENT

To understand the urgency of the evidence gap, consider the scale of deployment. Within months of ChatGPT's public release in November 2022, millions of students were using LLM-based tools for academic work. Universities scrambled to develop policies. Ed-tech companies rushed to integrate LLMs into existing products. Institutional adoption outpaced evidence by orders of magnitude — there were more position papers about AI in education in 2023 than there were empirical studies.

This is not unprecedented. The history of educational technology is a history of adoption outpacing evidence. But the pace and scale of LLM adoption are unprecedented, and the potential for both benefit and harm is proportionally larger. A calculator deployed in a mathematics classroom affects mathematical learning. An LLM deployed across all subjects affects learning across all domains — writing, reasoning, research, creative production, problem-solving. The scope of potential impact makes the evidence gap more consequential than any previous educational technology deployment.

## 5.4 WHAT COGNITIVE SCIENCE PREDICTS

Even without direct evidence on LLM-based educational tools, we can reason from established principles. The predictions are not speculative — they follow from well-replicated findings in cognitive science, instructional design, and motivation research.

**The testing effect predicts that easy access to answers will reduce learning.** The testing effect demonstrates that the act of effortful retrieval is what strengthens memory — not the act of encountering correct information. When a student struggles to recall a fact and then succeeds, that effortful retrieval produces stronger learning than re-reading the same fact would. When a student asks an LLM for the answer instead of attempting retrieval, they bypass the cognitive process that produces learning. The convenience of instant answers is antithetical to the retrieval practice that L1-003 identified as "one of the most powerful learning tools available."

This is not a theoretical concern. Students already prefer less effortful learning strategies (rereading, highlighting) over more effortful ones (retrieval practice, self-explanation) — the metacognitive error identified by L1-002. LLMs provide an even lower-effort path to what feels like learning: ask a question, receive a fluent, confident answer, move on. The subjective experience may be satisfying. The learning is likely shallow.

**Productive failure research predicts that immediate help will reduce conceptual understanding.** Kapur's (2024) research demonstrates that struggling with problems before receiving instruction produces deeper conceptual understanding and better transfer than instruction followed by practice. The mechanism is that struggle activates prior knowledge, reveals misconceptions, and creates a cognitive framework that makes subsequent instruction more meaningful. When students can immediately ask an LLM for help at the first sign of difficulty, they bypass this productive struggle. The LLM provides a clear, well-organized answer that feels like understanding but may actually prevent the deep processing that generates it.

This is precisely the risk Kapur himself identified in the foreword to *Productive Failure*: that generative AI could "tempt educators to skip the essential struggle that makes learning meaningful." The irony is that the better LLMs are at explaining — the more fluent, clear, and patient their responses — the more effective they may be at *preventing* the cognitive work that produces genuine understanding.

**Cognitive load theory predicts mixed effects depending on design.** On one hand, LLMs could reduce extraneous cognitive load by providing well-organized, clearly written explanations tailored to the learner's level — eliminating the need to search through textbooks, navigate poorly designed websites, or decode unclear prose. On the other hand, the conversational interface itself imposes cognitive demands: formulating questions, evaluating the reliability of responses, managing the interaction. For novice learners who lack the schemas to evaluate LLM output, the metacognitive load of determining whether an answer is correct may exceed the load of learning the material through traditional means.

**Self-determination theory predicts that AI scaffolding may undermine autonomy.** L1-002 established that autonomy — the sense of being the origin of one's actions — is a basic psychological need whose frustration undermines intrinsic motivation. Adaptive learning systems that prescribe learning paths, select problems, and provide solutions may support competence (by ensuring appropriate challenge levels) while undermining autonomy (by removing learner choice and control). Darvishi et al.'s (2023) finding that AI assistance reduced student agency is consistent with this prediction.

The tension is real: personalization requires the system to make decisions about what the learner should do next, which inherently reduces the learner's sense of autonomy. L1-004 resolved an analogous tension by showing that explicit instruction can be delivered autonomy-supportively — through providing rationale, offering choice within structure, and using invitational language. Whether AI systems can achieve the same balance is an open question with essentially no empirical evidence.

**Self-regulation research predicts differential effects by learner capacity.** L1-002 found that self-regulation can be taught but varies significantly across learners. Students with strong self-regulatory skills — the ability to set goals, monitor progress, choose effective strategies, and manage time — are likely to use AI tools more effectively than students without these skills. A self-regulated learner might use an LLM as a Socratic partner, testing their own understanding by explaining concepts to the AI and then evaluating its responses. A poorly self-regulated learner might use the same tool as an answer machine, bypassing cognitive effort entirely.

This predicted differential has profound equity implications. If AI tools primarily benefit students who already have strong self-regulation — which, as L1-002 established, correlates with socioeconomic background, educational privilege, and prior academic success — then AI in education could widen rather than narrow achievement gaps. The students who most need help learning would be least likely to use AI in ways that produce learning. This is the same pattern that has plagued every previous educational technology: tools designed to democratize learning end up benefiting those who are already advantaged.

**The generation effect predicts that producing is better than receiving.** A well-established finding in cognitive psychology holds that information you generate yourself is better retained than information you receive from an external source — the generation effect (Slamecka & Graf, 1978). When a student formulates their own explanation of a concept, even an imperfect one, they learn more than when they read or hear a perfect explanation. LLMs provide perfect explanations on demand, which may paradoxically reduce learning by eliminating the need for students to generate their own imperfect but more memorable versions.

This connects to the broader principle that learning is an active, effortful, generative process — not a receptive one. Technology that makes the reception of information effortless is optimizing for the wrong thing. The cognitive work of selecting, organizing, and integrating information — what Mayer identifies as the core of meaningful learning — is precisely the work that efficient AI tools are designed to eliminate. The more helpful the AI, the less cognitive work the student does, and the less the student learns.

## 5.5 THE COMPETENCE STACK ANALYSIS

Evaluating LLM-based educational tools against the five-layer competence stack:

**Layer 1 — Domain Knowledge:** LLMs can deliver content, explain concepts, generate practice questions, and provide immediate corrective feedback on factual recall. This is where they are most capable and where the risk is lowest — *provided* the technology is used for initial learning and not as a substitute for retrieval practice. The testing effect is clear: encountering correct information produces less learning than retrieving it. LLMs are excellent for initial exposure and explanation; they become counterproductive if they replace the effortful retrieval that consolidates knowledge.

A significant caveat is accuracy. LLMs produce confident, fluent output that may be factually incorrect — the "hallucination" problem. For domain knowledge at the novice level, learners lack the expertise to detect errors. This is a fundamental pedagogical problem: the tool that delivers content can also deliver misinformation, and the students who most need accurate content are least equipped to identify inaccuracy.

**Layer 2 — Skill:** Skill development requires deliberate practice with feedback — performing the skill, receiving correction, and performing again. LLMs can support this for some skills: writing (providing feedback on drafts), programming (identifying bugs, suggesting improvements), mathematical reasoning (checking solution steps). For physical skills, procedural skills requiring real-world interaction, or skills requiring tacit knowledge, LLMs are inapplicable.

The quality of LLM feedback for skill development is an open question. Jacobsen and Weber (2025) examined LLM-generated feedback and found that prompt engineering significantly affected quality, with well-designed prompts producing feedback that approached human quality in some dimensions but remained inferior in others — particularly in identifying the *root cause* of errors rather than simply flagging them. This connects to VanLehn's finding that effective tutoring requires engagement with the learner's reasoning process, not just their final output.

**Layer 3 — Judgment:** Judgment develops through exposure to varied, ambiguous situations in high-validity, high-feedback environments (Kahneman & Klein, 2009, as cited in L1-009). Can LLMs create such environments? Theoretically, they could generate case studies with ambiguous features, present scenarios requiring weighing competing considerations, and provide feedback on reasoning processes. But the design challenge is immense. Judgment development requires that the learner experience consequences of their decisions — that wrong judgment calls produce visible, meaningful outcomes. LLMs can simulate scenarios, but whether simulated consequences produce the same learning as real ones is unknown.

More fundamentally, LLMs are pattern-matching systems trained on text. They can model what competent judgment *sounds like* based on training data, but they do not exercise judgment themselves. When a learner asks an LLM whether their reasoning about a complex case is sound, the LLM produces a response that reflects the statistical distribution of how experts in the training data discuss such cases — not an evaluation of the learner's actual reasoning quality. This distinction matters. A human expert evaluating a learner's judgment brings domain expertise, contextual understanding, and genuine evaluative capacity. An LLM brings fluent pattern completion.

**Layer 4 — Metacognition:** This is where the interaction becomes most interesting and most dangerous. LLMs could theoretically support metacognition through prediction-first prompting ("Before I explain, what do you think the answer is?"), confidence calibration ("How confident are you in that answer?"), self-explanation prompting ("Can you explain why you chose that approach?"), and error analysis ("Let's look at where your reasoning went wrong"). These are exactly the techniques that L1-009 identified as effective for metacognitive development.

But there is a structural problem. Metacognitive development requires that the learner confront the limits of their own understanding — the uncomfortable experience of realizing they don't know what they thought they knew. This is what Kapur's productive failure achieves: the learner's initial failure makes their knowledge gaps visible *to themselves*. An LLM that immediately provides correct answers or detailed explanations when asked bypasses this confrontation. The learner never has to sit with their own confusion long enough to develop the metacognitive awareness that something is wrong.

The design implication is that effective LLM-based metacognitive support would need to *withhold* information strategically — to ask questions rather than provide answers, to prompt reflection rather than explain, to allow productive struggle rather than resolve difficulty. This is counterintuitive for a technology whose most impressive capability is generating fluent, helpful responses. The pedagogically optimal use of an LLM may be precisely the use that feels least helpful to the student.

**Layer 5 — Character and Disposition:** Intellectual honesty, tolerance for uncertainty, willingness to say "I don't know" — these epistemic virtues develop through environmental design, not through technology (L1-009). An LLM that confidently answers every question, never expresses uncertainty, and never says "I don't know" models exactly the wrong epistemic disposition. LLMs trained to be maximally helpful may be inadvertently modeling the performance of confidence that COMPETENCE-TARGET.md identifies as antithetical to genuine competence.

A more subtle risk: students who routinely use LLMs for academic work may develop the habit of producing polished output without genuine understanding — the "performance of competence" without actual competence.[1] This is the layer 5 equivalent of the automation complacency literature: when the system produces high-quality outputs, the human in the loop gradually loses the ability (and the inclination) to evaluate those outputs critically.

[1] The irony of this observation appearing in a paper produced by AI agents under human direction is not lost on us. We take it as a standing challenge: this lab must demonstrate that its principal investigator is developing genuine understanding of the material, not merely curating polished output. See the verification framework at LAB.APPLIEDPEDAGOGY.COM for how we attempt to hold ourselves to this standard.

## 5.6 THE OVER-RELIANCE RISK

The automation complacency literature from aviation, medicine, and industrial safety provides a framework for understanding the risks of AI-assisted learning. Wiener and Curry (1980) documented how flight-deck automation, while reducing certain categories of error, introduced new failure modes: pilots who trusted automated systems uncritically, failed to monitor automated processes adequately, and lost manual flying skills through disuse. Goddard, Roudsari, and Wyatt (2011) found similar patterns in clinical decision support systems: automation bias — the tendency to accept automated suggestions uncritically — was pervasive and resistant to training interventions.

Parker and Grote (2019) synthesized the automation literature broadly, finding that automation consistently produces a tradeoff: it reduces errors on routine tasks while degrading the human operator's ability to handle non-routine situations. The mechanism is skill atrophy through disuse combined with reduced situational awareness from over-trust in the automated system.

The parallels to education are direct. In aviation, automation was introduced to reduce pilot workload and prevent the human errors that caused crashes. In education, AI is being introduced to reduce teacher workload and prevent the instructional failures that produce poor learning outcomes. In both cases, the automation succeeds at its stated goal — fewer pilot errors on routine flights, more efficient content delivery and assessment — while creating a new category of risk: the degradation of the human capabilities that the automation was designed to supplement.

Parasuraman and Manzey (2010), in a widely cited review of automation complacency and bias, identified two distinct mechanisms: *complacency* (reduced monitoring of automated processes) and *automation bias* (the tendency to follow automated recommendations even when they are clearly wrong). Both mechanisms have been documented across domains — aviation, medicine, military operations, industrial process control — and both are relevant to AI-assisted learning.

Complacency in education would manifest as reduced self-monitoring: students who trust AI outputs stop checking them, stop evaluating whether the AI-generated explanation actually makes sense, stop noticing when the AI has made an error. Automation bias would manifest as students accepting AI-generated content uncritically, even when it conflicts with their own knowledge or when red flags are present. Both mechanisms would undermine the metacognitive skills that L1-009 identified as essential for full-stack competence.

Applied to education, the prediction is clear. Students who routinely rely on AI for writing, problem-solving, or reasoning may:

1. **Lose foundational skills** that they never fully develop or that atrophy through disuse. A student who asks an LLM to write first drafts may never develop the generative capacity that comes from staring at a blank page and producing ideas independently.

2. **Develop miscalibrated confidence.** Producing polished AI-assisted output may create the impression of competence where competence does not exist. The student knows how to *get* a good answer; they do not know how to *produce* one.

3. **Fail when the system fails.** When AI assistance is unavailable — in high-stakes professional situations, in novel contexts outside training data, in domains where AI generates plausible but incorrect output — students who have relied on AI may lack the independent capability to function.

4. **Never develop upper-stack competencies.** If AI assistance prevents the productive struggle that develops judgment, metacognition, and intellectual character, students may develop

extensive content knowledge (layer 1) and some procedural skill (layer 2) while remaining fundamentally underdeveloped in layers 3–5.

The evidence for these predictions in educational contexts is essentially zero — the technology is too new. But the predictions are not speculative; they follow from well-established findings in automation complacency research applied to a new domain.

## 5.7 PROMPT ENGINEERING AND TUTORING DESIGN

A nascent literature is beginning to examine how the *design* of AI tutoring interactions affects learning. Jacobsen and Weber (2025) examined the effect of prompt engineering on LLM feedback quality and found that well-designed prompts significantly improved the quality of AI-generated feedback — in some dimensions approaching the quality of human expert feedback. But the AI remained inferior at identifying the *root cause* of student errors. It could flag that an answer was wrong and suggest corrections, but it struggled to diagnose the underlying misconception that produced the error — exactly the “bug analysis” capability that made the most effective ITS successful.

Steinert et al. (2024) demonstrated that LLMs could be designed to provide research-based formative feedback — feedback grounded in learning science principles rather than generic chatbot responses. Their work showed that incorporating learning science into prompt design improved both the accuracy and the pedagogical quality of AI feedback. This is an important proof of concept: LLMs are not inherently pedagogically naive — they can be designed to implement learning science principles. But the design effort required is substantial, and most current deployments do not incorporate it.

The implications for Applied Pedagogy are clear. If LLM-based tools are used, their design must be informed by learning science — specifically by the feedback design principles established in L1-003 (task-focused, process-focused, actionable), the expertise-adaptive principles established in L1-004 (scaffold for novices, fade for intermediates, release for advanced learners), and the metacognitive training principles established in L1-009 (prediction-first, confidence calibration, self-explanation). Generic chatbot interactions — “How can I help you?” followed by an explanation — implement none of these principles and should be treated as the equivalent of problem-level ITS: easy to build, pedagogically shallow.

## 5.8 WHAT WOULD EFFECTIVE AI TUTORING LOOK LIKE?

Drawing on the ITS literature, Kapur’s productive failure research, and the metacognition literature from L1-009, an effective AI tutoring interaction might look something like this:

The learner encounters a challenging problem. Before any AI assistance is available, they must attempt the problem independently — a productive failure phase. The AI monitors their work at the step level (VanLehn’s inner loop), but does not intervene unless the learner requests help or reaches an impasse. When the learner does seek help, the AI does not provide the answer. Instead, it asks a Socratic question: “What approach have you tried so far?” or “What do you think would happen if you tried X?” If the learner provides an explanation, the AI evaluates the reasoning process — not just the answer — and provides process-focused feedback. If the learner’s confidence is miscalibrated (they are confident but wrong, or uncertain but right), the AI points this out explicitly as a metacognitive training moment.

This interaction design incorporates productive failure (Kapur), step-level feedback (VanLehn), process-focused feedback (L1-003), Socratic questioning (Nye et al.), confidence calibration (L1-

009), and expertise-adaptive scaffolding (L1-004). It is also much harder to build than a standard chatbot and much less satisfying to use — the student who wants a quick answer will find it frustrating. But the frustration is pedagogically productive, and the design is grounded in decades of evidence about what makes tutoring effective.

No such system has been rigorously evaluated. The description above is an extrapolation from established principles, not an evidence-based recommendation. But it illustrates the gap between what current LLM-based educational tools offer (answer-giving) and what the learning science literature says effective tutoring should look like (a carefully designed interaction that develops the learner's cognitive and metacognitive processes).

Part III

# FAILURE MODES AND SYNTHESIS

# 6 WHEN TECHNOLOGY HURTS LEARNING

## 6.1 THE DIGITAL NATIVE MYTH

Kirschner and De Bruyckere (2017) systematically debunked two popular beliefs: that today's students are "digital natives" who intuitively understand how to learn with technology, and that these students are skilled "multitaskers" who can effectively divide attention across multiple screens and tasks.

The evidence shows neither claim is true. Students who grew up with technology are comfortable *using* technology socially, but this does not translate to effective *learning* with technology. Being able to navigate Instagram is not the same as being able to evaluate the credibility of online sources, manage the cognitive demands of multimedia learning, or resist the temptation to multitask during instruction. The myth of the digital native is dangerous because it provides a justification for deploying technology in educational contexts without providing the scaffolding, instruction, and support that students actually need to learn with it.

## 6.2 DISTRACTION AND MULTITASKING

The evidence on classroom devices is concerning. Sana, Weston, and Cepeda (2013) demonstrated that laptop use during lectures reduced learning not only for the laptop user but for students sitting nearby who could see the screen. The effect operated through distraction — even when students intended to use laptops for note-taking, the availability of the internet proved irresistible, and the visual distraction affected their neighbors.

Wilmer, Sherman, and Chein (2017) reviewed the broader literature on mobile technology and cognition, finding consistent associations between heavy smartphone use and poorer attention, memory, and executive function. May and Elder (2018) reviewed media multitasking specifically in educational contexts and found negative associations with academic performance across multiple studies and methodologies.

The mechanism is straightforward: attentional switching between tasks (checking a notification, then returning to the lecture) incurs cognitive costs. Each switch requires reorienting working memory to the new task, and the residual activation from the previous task creates interference. For learning — which requires sustained, focused cognitive processing — these switching costs accumulate into meaningful learning deficits.

Vedechkina and Borgonovi (2021) reviewed the evidence on digital technology's effects on children's attention and cognitive control, finding that while the evidence is correlational and the effects are complex, the overall pattern suggests that heavy digital technology use is associated with poorer attention regulation — exactly the cognitive capacity most needed for effective learning.

## 6.3 THE ENGAGEMENT TRAP

Gamification in education illustrates the engagement trap. Sailer and Homner (2019) conducted a meta-analysis of gamification in learning and found that gamification produced small positive effects on cognitive learning outcomes ($d \approx 0.36$) and motivational outcomes. But the analysis revealed significant heterogeneity — some gamification approaches were effective and others were not, depending on the specific game elements used and how they were implemented.

The critical distinction is between gamification elements that support learning processes and those that merely increase time-on-task. Points, badges, and leaderboards — the most commonly implemented gamification elements — are extrinsic reward systems. L1-002 established that tangible, expected, contingent rewards reliably undermine intrinsic motivation (Deci et al., 1999). Gamification elements that take this form are likely to increase engagement in the short term while eroding the intrinsic interest that sustains long-term learning.

Dichev and Dicheva (2017) reviewed the gamification literature and concluded that "what is known" about gamification in education is far less than "what is believed." Many claims about gamification's benefits are not supported by rigorous evidence. Koivisto and Hamari (2019) similarly found that the field suffers from methodological weaknesses, including a reliance on short-duration studies that cannot capture the decay of novelty effects.

The engagement trap operates at the institutional level as well. Educational technology companies optimize for engagement metrics — time-on-task, session frequency, completion rates — because these are measurable and marketable. But engagement is not learning. A student who spends an hour on a gamified learning app collecting badges may learn less than a student who spends twenty minutes doing effortful retrieval practice with a stack of index cards. The metrics that drive the ed-tech industry are not the metrics that measure learning.

The disconnect between engagement metrics and learning outcomes is not accidental. It is a structural feature of the educational technology market. Engagement metrics are available in real time, scale across millions of users, and can be optimized through A/B testing. Learning outcomes require delayed post-tests with transfer items, control groups, and longitudinal follow-up — expensive, slow, and methodologically challenging. The market selects for what can be measured and marketed quickly, not for what matters educationally.

This creates a perverse incentive structure. An ed-tech product that is genuinely effective at producing learning — through spaced retrieval practice, productive failure, and effortful processing — will feel harder and less engaging to users than a product that optimizes for subjective enjoyment. The effective product will have lower engagement metrics, lower student satisfaction scores, and lower adoption rates in a market that evaluates products by these metrics. The market thus systematically selects for pedagogically inferior products.

Applied Pedagogy must resist this logic. When evaluating educational technology, the question is never "Do students like it?" or "Do they use it frequently?" but "Do they learn more, and does the learning transfer?" The answer to these questions requires rigorous evaluation designs that the ed-tech industry has little incentive to implement.

## 6.4 THE NOTE-TAKING QUESTION

The laptop-in-classroom debate intersects with the note-taking literature in illuminating ways. Mueller and Oppenheimer (2014) found that students who took notes by hand performed better on conceptual questions than students who took notes on laptops, even when the laptop students did not multitask. The proposed mechanism is that handwriting is slower, forcing students to process and summarize information as they write, while typing enables verbatim transcription — capturing the words without processing the meaning. This is another instance of desirable difficulty: the harder, slower method produces deeper processing and better learning.

The finding has been debated — some subsequent studies have found smaller or null effects, particularly when students are explicitly instructed not to transcribe verbatim. But the underlying principle is consistent with cognitive science: encoding quality depends on the depth of processing, and any tool that makes capture effortless risks reducing the cognitive engagement that produces learning. This principle extends beyond note-taking to any AI-assisted activity. When technology makes production easy — generating text, solving problems, producing code — it may reduce the cognitive effort that makes the production process a learning experience.

## 6.5 EQUITY AND ACCESS

Technology-dependent curricula assume reliable internet access, adequate devices, digital literacy, and a quiet space to work. These assumptions exclude millions of learners. The COVID-19 pandemic made this visible at scale: students without adequate internet access, devices, or home environments conducive to online learning were systematically disadvantaged when education moved online.

The equity dimension is not merely about access to hardware. It includes:

- **Digital literacy.** Students from families where parents use technology professionally arrive at school with different digital skills than students from families where technology use is limited to social media and entertainment. This is not a "digital native" gap; it is a socioeconomic gap expressed through technology.
- **Support structures.** Technology-mediated learning often requires more self-regulation, not less, because the social scaffolding of the classroom is absent. Students with strong self-regulation skills (which L1-002 found can be taught but often aren't) can thrive in technology-rich environments. Students without these skills may flounder.
- **Design bias.** Educational technology is typically designed by and for English-speaking, middle-class, Western users. The personalization algorithms, the cultural references, the assumptions about prior knowledge — all reflect the designers' context, not the diverse contexts of actual learners.

## 6.6 ADAPTIVE LEARNING PLATFORMS: THE UNFULFILLED PROMISE

Between ITS and LLMs, a generation of "adaptive learning platforms" emerged — products like Knewton, ALEKS, DreamBox, and Smart Sparrow — promising to personalize learning at scale. These platforms use algorithms to adjust content difficulty, pacing, and sequencing based on student performance data. They represent the most commercially successful category of educational AI, deployed in thousands of schools and reaching millions of students.

The evidence base for adaptive learning platforms is surprisingly thin relative to their market penetration. Fontaine et al. (2019) conducted a systematic review and meta-analysis of adaptive e-learning in health professions education and found small positive effects, but with significant methodological limitations in the included studies. Many evaluations were conducted or funded by the platform companies themselves — a conflict of interest that the broader education research community has not sufficiently scrutinized.

The fundamental problem with most adaptive learning platforms is that they adapt the *difficulty* and *pacing* of content delivery but not the *pedagogy*. They serve up easier problems when students struggle and harder problems when students succeed — a form of mastery learning that has some evidence base (L1-004). But they do not adapt the *type* of instruction based on what the learner needs. They do not shift from worked examples to productive failure as the learner develops expertise. They do not provide Socratic questioning when the learner has a misconception versus direct instruction when the learner lacks foundational knowledge. They personalize the *what* but not the *how* — and the *how* is where the learning science has the most to say.

This is the expertise reversal problem (L1-004) at platform scale. A system that delivers worked examples to a novice is doing the right thing; the same system delivering worked examples to an intermediate learner is imposing redundant cognitive load. Most adaptive platforms lack the sophistication to implement expertise-adaptive instruction — they adapt difficulty but not instructional approach. The result is a system that is modestly more efficient than one-size-fits-all instruction but fundamentally limited in its pedagogical impact.

## 6.7 DATA PRIVACY AND SURVEILLANCE

Educational technology creates surveillance infrastructure. Learning management systems, adaptive learning platforms, and AI tutoring tools collect granular data on student behavior: what they read, how long they spend on each page, what mistakes they make, when they work, where they pause. This data can be used for legitimate pedagogical purposes — identifying struggling students, adapting instruction, evaluating materials. But it also creates a surveillance environment that may undermine the psychological safety that L1-009 identified as essential for competence formation at layers 3–5.

When students know that every keystroke is being monitored, they may optimize for what the system measures rather than for genuine learning. This is Goodhart's Law applied to learning analytics: when the measure becomes the target, it ceases to be a good measure. The problem is not merely theoretical — it connects to the environmental multiplier that COMPETENCE-TARGET.md identifies as a first-order determinant of upper-layer competence.

L1-009 established that psychological safety — the ability to take risks, make errors, and express uncertainty without fear of punishment — is structurally necessary for competence formation at layers 3–5. Educational technology that monitors and evaluates student behavior in real time creates exactly the kind of surveillance environment that undermines psychological safety. A student who knows that their hesitation time, error rate, and strategy choices are being recorded and analyzed may become reluctant to take the risks — attempting hard problems, trying unconventional approaches, expressing genuine confusion — that are essential for learning.

The tension is real. The same data that could help teachers identify struggling students and provide targeted support can also create a surveillance environment that inhibits the risk-taking that learning requires. The resolution lies in design: data collection should be invisible to the learner during the learning process, used only for formative purposes, and never connected to evaluative consequences. Teachers should see aggregate patterns, not moment-by-moment keystroke data. And students should be explicitly told that the system is designed to help them learn, not to evaluate them — and the system's behavior should be consistent with this claim. A student who knows that time-on-task is being tracked may keep the application open without engaging with it. A student who knows that the system tracks error patterns may avoid challenging problems to maintain a high accuracy rate. The surveillance itself can undermine the risk-taking, error-making, and productive failure that effective learning requires.

# 7 THE COMPETENCE STACK ANALYSIS: WHAT CAN TECHNOLOGY ADDRESS?

This section evaluates technology's capacity at each layer of the competence stack. The analysis draws on all the evidence reviewed in previous chapters and the findings of L1-002 through L1-009. The purpose is not to provide a comprehensive technology evaluation but to give Applied Pedagogy a framework for deciding where technology can contribute to its mission and where it cannot.

## 7.1 LAYER 1: DOMAIN KNOWLEDGE — STRONG FIT

Technology excels at layer 1 knowledge delivery and is often superior to traditional methods for this specific purpose:

- **Spaced repetition software** implements proven cognitive science more efficiently than any human teacher could for factual knowledge.
- **Multimedia presentations** following Mayer's principles can present information more effectively than many live lectures.
- **ITS** can provide immediate corrective feedback on factual errors.
- **LLMs** can generate explanations, examples, and practice questions on demand.

The risk at layer 1 is not that technology is ineffective but that it is *too* effective at making knowledge acquisition feel easy. When the system provides clear explanations and immediate answers, it can bypass the effortful processing that produces durable learning. The design challenge is to use technology for initial exposure and explanation while ensuring that effortful retrieval practice — not passive consumption — drives long-term retention.

## 7.2 LAYER 2: SKILL — MODERATE FIT

Technology can support skill development in domains where practice is well-structured and feedback can be algorithmically assessed:

- **Programming environments** (with unit tests and linters) provide immediate feedback on code correctness.
- **Language learning apps** can assess pronunciation, grammar, and vocabulary use.
- **Mathematics tutoring systems** can evaluate solution steps and provide targeted feedback.
- **Writing assistants** can identify surface-level errors in grammar and mechanics.

Technology struggles with skills that require tacit knowledge, physical execution, interpersonal interaction, or judgment about quality in ill-structured domains. An ITS can teach a student to solve a system of equations; it cannot teach them to write a persuasive essay, conduct a difficult conversation, or diagnose a patient's illness from ambiguous symptoms.

### 7.3 LAYER 3: JUDGMENT — POOR FIT, WITH POSSIBILITIES

Judgment development requires varied exposure to ambiguous situations with meaningful consequences. Technology could theoretically support this through:

- **Case-based simulations** presenting ambiguous scenarios requiring weighing competing considerations.
- **Interactive fiction** presenting consequential decisions with delayed and uncertain outcomes.
- **AI-generated scenarios** that adapt to the learner's reasoning patterns.

But judgment development also requires a high-validity feedback environment where the learner can connect their decisions to outcomes. Most educational technology environments are low-validity — the consequences are artificial, the feedback is immediate and explicit, and the ambiguity is carefully controlled. Real judgment develops in environments where consequences are real, feedback is delayed and noisy, and the problem structure is genuinely uncertain. No current technology reproduces these conditions effectively.

### 7.4 LAYER 4: METACOGNITION — MIXED FIT

Technology has both promising and dangerous implications for metacognition:

**Promising:** Prediction-first interfaces that ask "What do you think will happen?" before showing results. Confidence calibration tools that track the accuracy of self-assessments over time. Self-explanation prompts built into learning materials. Analytics dashboards that make learning progress visible to learners.

**Dangerous:** Any technology that reduces the need for self-monitoring by providing external monitoring. Any system that tells students what they should study next, removing the metacognitive task of evaluating one's own knowledge gaps. Any tool that provides answers before the learner has had time to assess their own understanding. The risk is that technology externalizes metacognition — moves it from the learner's mind to the system — so that the learner never develops the internal capacity to monitor and regulate their own learning.

### 7.5 THE CROSS-LAYER PROBLEM

The layers of the competence stack do not operate independently. L1-009 found that they interact — knowledge enables judgment, metacognition enables skill development, and the environment shapes everything. This creates a problem for technology-based education: if technology effectively addresses layers 1–2 while failing at layers 3–5, it does not simply produce partial competence — it may produce a *specific kind* of incompetence characterized by extensive content knowledge without the judgment, metacognitive awareness, or epistemic character to deploy it wisely.

Consider the medical analogy. A medical student who uses AI to master vast amounts of diagnostic knowledge (layer 1) and even procedural skill (layer 2) but never develops clinical judgment (layer 3), metacognitive awareness of their own diagnostic limitations (layer 4), or the intellectual honesty to acknowledge uncertainty (layer 5) is not merely "partially competent." They are specifically dangerous — their impressive knowledge base creates confidence that their underdeveloped upper layers cannot justify. The competence stack is not a ladder where each rung is independently valuable; it is an integrated system where the upper layers provide the context that makes the lower layers useful.

This is the deepest concern about technology-dominated education: it may produce people who *know* a great deal but *understand* little, who can *produce* impressive outputs but cannot *evaluate* them, who appear competent on standardized assessments but fail in situations requiring judgment, self-awareness, or intellectual courage. Applied Pedagogy's competence stack was designed to identify exactly this failure mode. Technology that addresses only the bottom of the stack is not merely incomplete — it is potentially competence-distorting.

## 7.6 LAYER 5: CHARACTER AND DISPOSITION — NO DIRECT FIT

Technology cannot develop intellectual honesty, tolerance for uncertainty, or the courage to say "I don't know." These dispositions are environmentally shaped (L1-009). Technology can, however, create environments that are more or less conducive to their development:

- **Negative:** AI systems that model confident certainty on every topic. Gamification that rewards speed and correctness over careful thought. Analytics that create performance pressure.
- **Positive:** Systems designed to normalize uncertainty — "This is a hard problem with no clear answer" rather than "The answer is…" Discussion forums that reward thoughtful disagreement. Feedback systems that celebrate accurate self-assessment ("You correctly identified that you didn't know this") as much as correct answers.

The design of the technology environment is a layer 5 intervention, even if the technology itself cannot directly train epistemic character.

There is one specific way that LLMs could contribute negatively at layer 5 that deserves emphasis. LLMs are trained to be helpful, informative, and confident. They rarely express genuine uncertainty — they may qualify statements with hedging language, but the overall tenor is one of assured expertise. A student who spends hours interacting with a system that always has an answer, always sounds confident, and never says "I genuinely don't know" is being implicitly taught that competence looks like perpetual confidence. This is the opposite of the epistemic character that COMPETENCE-TARGET.md identifies as essential: the willingness to sit with uncertainty, to say "I don't know," to resist the pressure to perform confidence when knowledge is absent.

The remedy, if LLMs are to be used educationally, is to design systems that deliberately model epistemic humility — that sometimes say "This is a genuinely hard question and I'm not confident in my answer," that sometimes refuse to answer ("I could give you an answer, but it might not be accurate — let's think through this together"), and that consistently frame knowledge as provisional rather than certain. Whether such systems would be commercially viable is doubtful. Whether they would be pedagogically valuable is, by the logic of the competence stack, clear.

# 8

# WHAT THE CONNECTIONS TO OTHER L1 INVESTIGATIONS REVEAL

The findings of this investigation do not stand alone. They intersect with the conclusions of every completed L1 agent in ways that strengthen the overall picture and sharpen the prescriptions.

**Connection to L1-002 (Motivation).** The motivation investigation established that extrinsic rewards reliably undermine intrinsic motivation (Deci et al., 1999), that autonomy support is essential for sustained engagement, and that self-regulation is teachable but requires explicit instruction. Educational technology intersects with each finding. Gamification elements that function as extrinsic rewards — points, badges, leaderboards — risk the undermining effect. Adaptive systems that prescribe learning paths may undermine autonomy. AI tools that externalize self-regulation (by telling students what to study and when) may prevent the development of the self-regulatory capacity that L1-002 identified as the "meta-skill that enables all other learning." The convergence is stark: the features that make educational technology *convenient* are often the features that make it *motivationally counterproductive*.

**Connection to L1-003 (Assessment).** The assessment investigation established that the testing effect is one of the most robust findings in cognitive psychology, that feedback should be task-focused and process-focused rather than person-focused, and that grades negate the benefit of feedback. Technology can implement both the testing effect (through SRS and automated low-stakes quizzing) and effective feedback (through well-designed automated responses). But it can also undermine both — LLMs that provide answers bypass retrieval practice, and AI-generated feedback that evaluates the person ("Great job!") rather than the work ("Your argument in paragraph three lacks supporting evidence") replicates the worst patterns of human feedback. The L1-003 principle that assessment should be "frequent, low-stakes, and formative" is easily implemented through technology; the principle that feedback should address reasoning processes requires far more sophisticated design.

**Connection to L1-004 (Instructional Design).** The instructional design investigation established the expertise continuum — novices need explicit instruction while developing learners benefit from productive failure and guided inquiry. This has direct implications for educational technology design: a system that provides the same level of scaffolding regardless of learner expertise is violating the expertise reversal effect. Most adaptive learning platforms adjust difficulty but not instructional approach, missing the crucial insight that the *type* of instruction should change as the learner develops, not just its *level*. An effective AI tutoring system would need to detect not just whether the learner got the answer right, but where the learner sits on the expertise continuum, and adjust its interaction style accordingly — from worked examples for novices to Socratic questioning for intermediates to minimal intervention for advanced learners.

**Connection to L1-009 (Competence Formation).** The competence formation investigation established that the learning environment is a first-order intervention, that feedback loops are the infrastructure of competence development, and that upper-layer competencies (judgment, metacognition, character) are the most consequential and the least addressable through traditional instruction. Technology's relationship to each finding is ambivalent. Technology can create environments that support or undermine psychological safety. Technology can shorten feedback loops (immediate automated feedback) or lengthen them (grades delivered days after submission). Technology primarily addresses lower-layer competencies while the lab's distinctive contribution lies at the upper layers. The L1-009 finding that "environment first" should be the design priority directly challenges the technology-first approach that dominates ed-tech discourse.

The convergence across all five investigations points to a unified conclusion: educational technology is most useful when it implements established learning science principles more efficiently than human teachers can (spaced repetition, immediate feedback on well-structured tasks, adaptive pacing), and most dangerous when it undermines the motivational, metacognitive, and environmental conditions that produce genuine competence. The technology is a means, not an end. The learning science is the end.

# Part IV

# PRESCRIPTIONS

# 9 DESIGN PRINCIPLES FOR APPLIED PEDAGOGY

From the evidence reviewed, the following principles should guide Applied Pedagogy's use of educational technology. They are ordered from most strongly evidence-based to most extrapolated from principles.

## 9.1 PRINCIPLE 1: TECHNOLOGY IS A DELIVERY MECHANISM, NOT A PEDAGOGY

The same learning science principles apply regardless of medium. Technology can make some principles easier to implement (spaced repetition, immediate feedback, adaptive pacing) but cannot substitute for them. Evaluate every technology tool against established principles of cognitive load management, retrieval practice, feedback design, autonomy support, and expertise-adaptive instruction — not against engagement metrics or marketing claims.

## 9.2 PRINCIPLE 2: PROTECT PRODUCTIVE STRUGGLE

Design AI-assisted learning to *preserve* the cognitive effort that produces learning, not to eliminate it. Specific implementations:

- **Delay AI assistance.** Require students to attempt problems independently before AI help becomes available. Implement productive failure sequences: struggle first, then AI-scaffolded resolution.
- **Use AI for Socratic questioning, not answer-giving.** Program AI interactions to ask questions, prompt self-explanation, and guide reasoning rather than providing solutions. "What do you think will happen?" and "Can you explain your reasoning?" are pedagogically superior to "Here is the answer."
- **Make retrieval practice the default.** Use technology to deliver retrieval practice (flashcards, quizzes, recall prompts) rather than content delivery as the primary learning activity.

## 9.3 PRINCIPLE 3: IMPLEMENT SPACED REPETITION FOR LAYER 1 KNOWLEDGE

Use SRS technology for factual knowledge acquisition — vocabulary, definitions, historical dates, scientific terminology, procedural steps. This is the highest-confidence application of educational technology, grounded in over a century of replicated findings. Design the SRS to incorporate:

- Free-recall and short-answer formats (more effortful, more effective than recognition).
- Immediate corrective feedback after each retrieval attempt.
- Interleaving across topics to prevent context-dependent learning.
- Integration with the broader curriculum so that SRS-reinforced knowledge is applied in meaningful contexts, not learned in isolation.

### 9.4 PRINCIPLE 4: APPLY MAYER'S PRINCIPLES TO ALL DIGITAL MATERIALS

Audit every piece of digital learning material against the multimedia learning principles. Specifically:

- Remove decorative elements that do not serve instructional objectives (coherence).
- Do not display text while narrating it (redundancy).
- Present corresponding words and pictures together, not on separate screens (spatial contiguity).
- Segment complex lessons into learner-paced chunks (segmenting).
- Provide pre-training on key concepts before complex multimedia presentations (pre-training).
- Use conversational language (personalization).

### 9.5 PRINCIPLE 5: DESIGN AI INTERACTIONS FOR METACOGNITION

If LLM-based tools are used, design them to *develop* metacognition rather than replace it:

- **Prediction-first prompting.** Before the AI provides information, ask the learner to predict, estimate, or hypothesize. This forces metacognitive engagement and creates the prediction-reality gap that L1-009 identified as a powerful metacognitive signal.
- **Confidence calibration.** Ask learners to rate their confidence in their answers before checking. Track calibration accuracy over time. Celebrate accurate uncertainty.
- **Self-explanation prompting.** After the learner provides an answer, ask them to explain their reasoning before providing feedback. This promotes the self-explanation effect (Chi et al., 1989, as cited in L1-009).
- **Strategic withholding.** The AI should sometimes *not* answer — should say "I could tell you, but try reasoning through it first" or "What do you already know about this?" The pedagogically optimal AI tutor is one that talks less than the student.

### 9.6 PRINCIPLE 6: USE ITS PRINCIPLES FOR ANY AUTOMATED TUTORING

Decades of ITS research have established what effective automated tutoring looks like. Any LLM-based tutoring system should incorporate:

- **Step-level engagement** — interact with the learner during problem-solving, not just after.
- **Process-focused feedback** — address reasoning strategies, not just answer correctness.
- **Adaptive scaffolding** — provide more support for novices, less for developing learners, following L1-004's expertise continuum.
- **Mastery-based progression** — advance to new material only when current material is mastered.
- **Bug modeling** — identify specific misconceptions and address them directly.

### 9.7 PRINCIPLE 7: AUDIT TECHNOLOGY AGAINST THE FULL COMPETENCE STACK

Before adopting any educational technology, evaluate it against all five layers:

| Question | If "no," the technology is insufficient |
|---|---|
| Does it support knowledge acquisition? | Layer 1 gap |
| Does it enable deliberate practice with feedback? | Layer 2 gap |
| Does it expose learners to varied, ambiguous situations? | Layer 3 gap |
| Does it require learners to monitor their own understanding? | Layer 4 gap |
| Does the technology environment model epistemic virtues? | Layer 5 gap |

Most technologies will address only layers 1–2. This is acceptable if the technology is positioned as a component of a broader educational design that addresses all five layers — but it is unacceptable if the technology is positioned as the complete learning experience.

## 9.8 PRINCIPLE 8: GUARD AGAINST THE ENGAGEMENT TRAP

Distinguish engagement from learning. Specifically:

- Do not use time-on-task, session frequency, or completion rates as proxies for learning. These are engagement metrics, not learning metrics.
- Be skeptical of gamification elements that function as extrinsic rewards (points, badges, leaderboards). These may increase engagement while undermining intrinsic motivation.
- Evaluate technology tools by their effect on learning outcomes (measured through delayed post-tests with transfer items), not by student satisfaction surveys, which consistently favor easier, less effective learning conditions.

## 9.9 PRINCIPLE 9: MINIMIZE SURVEILLANCE, MAXIMIZE SAFETY

Design the technology environment to support the psychological safety that upper-layer competence requires:

- Collect only the data needed for legitimate pedagogical purposes.
- Make data collection transparent to learners.
- Do not use learning analytics for surveillance or compliance monitoring.
- Create technology-mediated spaces where error is safe — where making mistakes is expected, visible, and informational rather than punitive.
- Ensure that AI-mediated assessment is low-stakes and formative, not high-stakes and summative.

## 9.10 PRINCIPLE 10: BUILD AI LITERACY INTO THE CURRICULUM

If students are going to use AI tools — and they will, regardless of institutional policy — they need explicit instruction in how to use them effectively. This is not "digital literacy" in the vague sense of knowing how to use computers. It is specific, evidence-informed instruction in:

- **The limits of AI accuracy.** Students need to understand that LLMs produce fluent, confident output that may be factually incorrect. They need practice detecting AI errors — a metacognitive skill that must be deliberately cultivated.
- **The cognitive costs of AI assistance.** Students need to understand the testing effect, productive failure, and the relationship between effort and learning. They need to know that using

AI to avoid cognitive effort is counterproductive for learning, even when it produces correct outputs.

- **Automation bias.** Students need to understand the documented tendency to accept automated suggestions uncritically, and they need strategies for maintaining critical evaluation of AI output.
- **The distinction between production and learning.** Students need to understand that producing a correct answer with AI assistance is not the same as learning to produce a correct answer independently. The production is a product; the learning is a process. AI can help with the product while undermining the process.

This instruction should be embedded in the curriculum, not delivered as a separate "AI literacy" module, following the same principle that L1-002 established for self-regulation instruction: teach it within the domain, model it, scaffold it, and repeat it across contexts.

## 9.11 PRINCIPLE 11: MAINTAIN HUMAN RELATIONSHIPS AS THE CORE

Technology should supplement, not replace, the human relationships that sustain learning. The L1-002 finding that relatedness is a basic psychological need, the L1-009 finding that modeling is essential for upper-layer competence, and the ITS finding that human tutoring remains superior for complex outcomes all converge on the same conclusion: the most important educational technology is a skilled, caring human teacher. Technology that reduces teacher-student interaction, replaces mentoring with algorithms, or substitutes AI conversation for human conversation is moving in the wrong direction regardless of its efficiency gains at layers 1–2.

# 10 CLOSING ASSESSMENT: CONFIDENCE LEVELS

## 10.1 WHAT WE KNOW WITH HIGH CONFIDENCE

- **ITS produce moderate learning gains** (d ≈ 0.66) compared to conventional instruction in well-structured domains, with step-level feedback being the critical feature (VanLehn, 2011; Kulik & Fletcher, 2016; Ma et al., 2014).
- **Mayer's multimedia principles** are well-replicated and provide reliable design guidance for digital learning materials (Mayer, 2002, 2020).
- **Spaced repetition software** effectively implements proven cognitive science for factual knowledge acquisition (Kornell, 2009; Settles & Meeder, 2016; Gilbert et al., 2023).
- **More technology can mean less learning** when technological novelty consumes cognitive resources without serving instructional objectives (Makransky et al., 2019; Makransky & Petersen, 2021).
- **Classroom devices cause distraction** that impairs learning for both users and nearby peers (Sana et al., 2013; Wilmer et al., 2017).
- **Gamification effects are small and heterogeneous,** with extrinsic reward elements risking motivational undermining (Sailer & Homner, 2019; L1-002).
- **Students are not "digital natives"** who intuitively know how to learn with technology (Kirschner & De Bruyckere, 2017).

## 10.2 WHAT WE KNOW WITH MODERATE CONFIDENCE

- **ITS are limited to well-structured domains** and primarily address layers 1–2 of the competence stack. Their effectiveness in ill-structured domains is unproven.
- **LLM accuracy is insufficient** for unsupervised use by novice learners who cannot detect hallucinations.
- **AI assistance may reduce student agency,** consistent with SDT predictions (Darvishi et al., 2023).
- **Automation complacency** from other domains predicts over-reliance risks in education (Goddard et al., 2011; Wiener & Curry, 1980).
- **Technology-dependent curricula** exacerbate equity gaps.

## 10.3 WHAT WE EXTRAPOLATE FROM PRINCIPLES (LOW CONFIDENCE, AWAITING DIRECT EVIDENCE)

- **LLMs may reduce productive struggle** by providing easy access to answers that bypass effortful cognitive processing.
- **LLMs may impair metacognitive development** by externalizing self-monitoring functions.
- **LLMs may undermine upper-stack competence** (layers 3–5) while supporting lower-stack learning (layers 1–2).
- **AI tutoring designed with Socratic, prediction-first, withholding principles** may be more effective than answer-giving designs.

- **The pedagogically optimal use of AI may feel unhelpful to students** because the most effective learning experiences involve struggle and difficulty, not ease and fluency.

## 10.4 THE HONEST BOTTOM LINE

Educational technology has a thirty-year track record of delivering modest improvements in knowledge acquisition and skill development through well-designed systems that implement established cognitive science principles. The ITS literature demonstrates that automated tutoring can approach — but not exceed — the effectiveness of human tutoring in well-structured domains. Mayer's principles provide reliable design guidance. Spaced repetition software genuinely works for its narrow target use case.

The LLM revolution adds a technology that is more flexible, more accessible, and more impressive than any previous educational technology. But it also brings unprecedented risks to the cognitive processes that produce genuine learning — risks that follow directly from established cognitive science but have not yet been empirically tested in educational contexts.

Applied Pedagogy's approach to technology should be neither uncritically enthusiastic nor reflexively Luddite. The evidence supports using technology for what it does well (content delivery, spaced retrieval, immediate feedback on well-structured tasks) while being vigilant about what it does poorly (developing judgment, supporting metacognition, building character) and what it might actively harm (productive struggle, learner autonomy, epistemic honesty).

## 10.5 THE HISTORICAL PERSPECTIVE

Every previous wave of educational technology has followed the same trajectory: exaggerated promise, enthusiastic adoption, disappointing evaluation, and eventual integration as one tool among many. Each wave has left behind genuine contributions — overhead projectors made diagrams possible, photocopiers enabled distributed practice materials, the internet democratized access to information, LMS platforms enabled formative assessment at scale — but none has transformed learning in the way its proponents predicted.

The reason is always the same: the technology changes the delivery mechanism, but the learning still happens in the human brain, subject to the same cognitive architecture, the same motivational dynamics, the same developmental constraints. The testing effect does not care whether retrieval practice is delivered via flashcard, computer screen, or AI conversation. Productive failure does not care whether the problem is presented on paper or in virtual reality. Autonomy support does not care whether the choice architecture is designed by a teacher or an algorithm.

LLMs are different from previous technologies in their scope and flexibility — they can generate natural language, reason about complex topics, and interact conversationally in ways that no previous technology could. But they are not different in the fundamental sense that matters: they are still delivery mechanisms for learning experiences, and the quality of those experiences is determined by the pedagogical design, not the technological capability.

## 10.6 THE RESPONSIBLE PATH FORWARD

Applied Pedagogy's approach to educational technology should be characterized by three commitments:

**Evidence over enthusiasm.** Every technology deployment should be treated as an hypothesis to be tested, not a solution to be implemented. The evidence base for LLM-based educational tools is essentially zero. Responsible deployment means gathering evidence — measuring learning outcomes with delayed post-tests and transfer items, not just engagement metrics and student satisfaction — and being willing to abandon tools that do not demonstrate learning gains.

**Principles over platforms.** The learning science principles established by decades of research — retrieval practice, productive failure, expertise-adaptive instruction, formative feedback, autonomy support, metacognitive training, psychological safety — are the constants. The technological platforms are variables. Design the learning experience around the principles, then select technology that supports the design — not the other way around.

**Full-stack accountability.** Any educational technology adopted by Applied Pedagogy must be evaluated against the full competence stack. A tool that demonstrably improves layer 1 knowledge acquisition but undermines layer 3 judgment development, layer 4 metacognitive capacity, or layer 5 epistemic character is a net negative for the mission, regardless of its impressiveness on narrow measures. The defining question is not "How can we use AI in education?" but "How can we use AI in ways that develop the full competence stack — not just the layers that technology finds easy?"

The answer to this question will unfold over years of careful experimentation, evaluation, and iteration. The honest conclusion of this investigation is that the learning science provides clear principles, the ITS literature provides documented lessons, the multimedia framework provides design constraints — and the direct evidence for how to use LLMs effectively in education is almost entirely missing. We know enough to avoid the most obvious mistakes. We do not yet know enough to confidently prescribe the optimal design. That humility — the willingness to say "we don't know yet" — is itself an expression of the epistemic character that the competence stack demands.

There is a final irony worth noting. The tool that this lab uses for research — an AI system capable of reading papers, synthesizing evidence, and generating prose — is itself an instance of the technology under investigation. The very system that produced this analysis could, in a different context, be the system that undermines a student's productive struggle with the same material. The difference lies in the design of the interaction, the purpose it serves, and the human judgment that governs its use. Technology is not the problem. Nor is it the solution. It is an amplifier — of effective pedagogy and of ineffective pedagogy alike. The learning science tells us what effective pedagogy looks like. Applied Pedagogy's task is to ensure that the amplifier is pointed in the right direction.

*Dissertation complete. L1-005 investigation. 38 sources consulted.*

# BIBLIOGRAPHY

Baker, R. S. (2016). Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education*, 26(2), 600–614.

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.

Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D., & Siemens, G. (2023). Impact of AI assistance on student agency. *Computers & Education*, 210, 104967.

Dichev, C., & Dicheva, D. (2017). Gamifying education: What is known, what is believed, and what remains uncertain. *International Journal of Educational Technology in Higher Education*, 14(1), 9.

Escueta, M., Nickow, A., Oreopoulos, P., & Quan, V. (2020). Upgrading education with technology: Insights from experimental research. *Journal of Economic Literature*, 58(4), 897–946.

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, 62(3), 460–474.

Gilbert, M. M., et al. (2023). A cohort study assessing the impact of Anki as a spaced repetition tool on academic performance in medical school. *Medical Science Educator*, 33, 955–962.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121–127.

Gordon, M., et al. (2024). A scoping review of artificial intelligence in medical education: BEME Guide No. 84. *Medical Teacher*, 46(4), 446–470.

Hamilton, D., McKechnie, J., Edgerton, E., & Wilson, C. (2020). Immersive virtual reality as a pedagogical tool in education: A systematic literature review. *Journal of Computers in Education*, 8(1), 1–32.

Holstein, K., McLaren, B. M., & Aleven, V. (2019). Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity. *Journal of Learning Analytics*, 6(2), 27–52.

Jacobsen, L. J., & Weber, K. E. (2025). The promises and pitfalls of large language models as feedback providers. *AI*, 6(2), 35.

Jape, D., Zhou, J., & Bullock, S. (2022). A spaced-repetition approach to enhance medical student learning and engagement in medical pharmacology. *BMC Medical Education*, 22, 337.

Jošt, G., Taneski, V., & Karakatič, S. (2024). The impact of large language models on programming education and student learning outcomes. *Applied Sciences*, 14(10), 4115.

Kapur, M. (2024). *Productive Failure: Unlocking Deeper Learning Through the Science of Failing*. Jossey-Bass.

Kasneci, E., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274.

Kirschner, P. A., & De Bruyckere, P. (2017). The myths of the digital native and the multitasker. *Teaching and Teacher Education*, 67, 135–142.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm. *Cognitive Science*, 36(5), 757–798.

Koivisto, J., & Hamari, J. (2019). The rise of motivational information systems: A review of gamification research. *International Journal of Information Management*, 45, 191–210.

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology*, 23(9), 1297–1317.

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

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, 13(4), 410.

Ma, W., Adesope, O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, 106(4), 901–918.

Makransky, G., & Mayer, R. E. (2022). Benefits of taking a virtual field trip in immersive virtual reality: Evidence for the immersion principle in multimedia learning. *Educational Psychology Review*, 34, 1771–1798.

Makransky, G., & Petersen, G. B. (2021). The Cognitive Affective Model of Immersive Learning (CAMIL). *Educational Psychology Review*, 33(3), 937–958.

Makransky, G., Terkildsen, T. S., & Mayer, R. E. (2019). Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learning and Instruction*, 60, 225–236.

May, K. E., & Elder, A. D. (2018). Efficient, helpful, or distracting? A literature review of media multitasking in relation to academic performance. *International Journal of Educational Technology in Higher Education*, 15(1), 13.

Mayer, R. E. (2002). *Multimedia Learning*. Cambridge University Press.

Mayer, R. E. (2020). *Multimedia Learning* (3rd ed.). Cambridge University Press.

Meyer, O. A., Omdahl, M. K., & Makransky, G. (2019). Investigating the effect of pre-training when learning through immersive virtual reality and video. *Computers & Education*, 140, 103603.

Molenaar, I. (2022). Towards hybrid human-AI learning technologies. *European Journal of Education*, 57(4), 632–645.

Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, 24(4), 427–469.

Parker, S. K., & Grote, G. (2020). Automation, algorithms, and beyond: Why work design matters more than ever in a digital world. *Applied Psychology*, 71(3), 1171–1204.

Popenici, S. A. D., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, 12(1), 22.

Roll, I., & Wylie, R. (2016). Evolution and revolution in artificial intelligence in education. *International Journal of Artificial Intelligence in Education*, 26(2), 582–599.

Sailer, M., & Homner, L. (2019). The gamification of learning: A meta-analysis. *Educational Psychology Review*, 32, 77–112.

Sana, F., Weston, T., & Cepeda, N. J. (2013). Laptop multitasking hinders classroom learning for both users and nearby peers. *Computers & Education*, 62, 24–31.

Settles, B., & Meeder, B. (2016). A trainable spaced repetition model for language learning. *Proceedings of ACL*, 1848–1858.

Steinert, S., et al. (2024). Harnessing large language models to develop research-based learning assistants for formative feedback. *Smart Learning Environments*, 11, 46.

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

Vedechkina, M., & Borgonovi, F. (2021). A review of evidence on the role of digital technology in shaping attention and cognitive control in children. *Frontiers in Psychology*, 12, 611155.

Wiener, E. L., & Curry, R. E. (1980). Flight-deck automation: Promises and problems. *Ergonomics*, 23(10), 995–1011.

Wilmer, H. H., Sherman, L. E., & Chein, J. M. (2017). Smartphones and cognition: A review of research exploring the links between mobile technology habits and cognitive functioning. *Frontiers in Psychology*, 8, 605.

Yan, L., et al. (2023). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(4), 1340–1373.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education. *International Journal of Educational Technology in Higher Education*, 16(1), 39.