THE STATE OF THE RESEARCH

What Applied Pedagogy Has Found, How Confident We Are, and What Comes Next

Applied Pedagogy Research Lab

Guido Bartolucci, Principal Investigator

guido@appliedpedagogy.com

LAB.APPLIEDPEDAGOGY.COM

Wave 1 Retrospective · April 2026

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

CONTENTS

Part I

WHY THIS WORK EXISTS

1

THE MOTIVATION

Genuinely competent people — the ones who notice what others miss, reason from first principles, update when they are wrong, and know what they do not know — appear to be exceptions. Not because most people lack the capacity, but because nothing in the standard educational apparatus is designed to develop it. The people we most admire for their judgment, their intellectual honesty, their ability to navigate situations nobody prepared them for — they arrived at those qualities through some combination of luck, temperament, and circumstances that no curriculum specified and no institution measured. Competence at the layers that matter most happens despite formal education, not because of it.

This is a staggering waste. The capability approach (Sen, Nussbaum) frames education as the development of capabilities to move through life — capabilities that change across the lifespan, both in what is required and in what a person becomes capable of acquiring. Education, through this lens, is not a sixteen-year front-loaded process that ends with a credential. It is a lifelong project of expanding what a person can do, decide, create, and become. When it works, it produces people who can handle what the world actually asks of them. When it fails — and it fails routinely — it produces people who stopped being curious at eleven, who cannot learn without being told to, who follow scripts written by someone else because they were never given the tools to write their own.

The failure is not inevitable. The science of how humans learn is real. It is decades old in many areas, well replicated, and remarkably consistent across independent research traditions. We know which study strategies work and which do not, and the ones that students are taught and shown by the culture around them — rereading, highlighting, cramming — are among the least effective ever evaluated. We know that extrinsic rewards undermine intrinsic motivation, and the entire grading system is an extrinsic reward apparatus. We know that judgment develops only through varied, consequential, ambiguous situations with structured feedback, and the standard classroom provides none of these. We know that environments which penalize honesty degrade the capacity to perceive truth over time, and the hidden curriculum of most institutions rewards confident performance over honest engagement with reality.

The science exists. The practice does not follow it. Education is in a state analogous to medicine before evidence-based practice — driven by tradition, authority, anecdote, and institutional inertia rather than by what the evidence actually supports. The infrastructure that medicine built to close its own science-practice gap — systematic reviews, clinical guidelines, institutional mandates — does not exist for education. Nearly half of teachers endorse neuromyths such as learning styles (Dekker et al., 2012). Textbooks lag the research by a decade. The curricula that shape how a billion children spend their formative years are assembled without explicit reference to what learning science has established about how learning works, what develops motivation, or what destroys it.

Meanwhile, the arrival of capable AI has created a triple crisis. It has broken traditional assessment — students can produce polished work without understanding it, exposing evaluation systems that never measured what they claimed to measure. It is restructuring white-collar work — the employment contract that justified credential acquisition for two generations is dissolving, and the people most vulnerable are those whose education developed Layers 1 and 2 (knowledge and skill) while neglecting Layers 3 through 5 (judgment, metacognition, character). And it is

exposing the shallowness of educational systems that optimized for information transfer in an era when information is free and infinite.

This is not an abstract concern. The author's children are four and six years from adulthood. The question of whether education can produce competent, capable, curious, and kind adults is not a research interest — it is a deadline.

But the same technology that is causing the crisis is also, for the first time, making it possible for a small team of concerned people to survey, cross-reference, and synthesize the research literature across ten domains in months rather than decades — and to build the tools that translate what the science says into something practitioners can use. That is what this lab is doing. The question this document answers is: what have we found so far?

2 THE ASSUMPTIONS

Every serious project starts from assumptions. Here are ours.

Human capability is largely wasted. A more capable person lives a richer life — they can navigate what the world actually asks of them, find meaning in contribution, and shape their circumstances rather than merely endure them. At the scale of a society, more capable people means harder problems solved, faster adaptation, and a world that actually improves.

Education is lifelong and self-directed. If an educational approach cannot produce a person who continues to learn after the instruction stops, it has failed at the thing that matters most.

Capabilities, not credentials, are the measure of success. Following Sen and Nussbaum: educational success is expanding what a person can do and become. Not what diploma they hold. What they can actually do when the world places a situation in front of them.

There are two crises and an opportunity. AI has broken traditional assessment and is dissolving the employment contract that justified credential acquisition for two generations. These are crises. The opportunity is that the same technology makes it possible to survey, synthesize, and cross-reference the research literature at a scale that was previously impossible — and to build tools that translate what the science says into something practitioners can use.

Incompetence is the anti-goal. The competence stack is most vivid in its absence — in the person who holds a consequential position but cannot identify the boundaries of their own knowledge, cannot receive bad news without punishing the messenger, and performs confidence instead of engaging with reality. A conventional education can produce this person and frequently does.

Competence is the aspiration. The "competence porn" tradition in fiction — characters who are compelling because they work the problem — captures something real. They notice, reason from first principles, update on evidence, and know what they don't know. In the real world, the Apollo program, Crew Resource Management, and blameless post-mortem cultures are what institutional competence looks like at scale.

3 THE PROBLEM WE ARE TRYING TO SOLVE

The specific problem is this: the science of learning exists, but no institution has the structural incentives to synthesize it honestly, evaluate it against a coherent framework of what competence means, and translate it into tools and curricula that people can use.

This is not a gap that requires new research. The primary research exists — thousands of papers, hundreds of meta-analyses, decades of converging evidence from cognitive science, motivational psychology, assessment theory, training science, and developmental psychology. What does not exist is the synthesis. The cognitive scientists do not read the training science literature. The motivation researchers do not read the institutional analysis. The teachers colleges teach a version of it, filtered through accreditation requirements and textbooks that lag the research by a decade.

Applied Pedagogy's contribution is the synthesis itself — reading across domains that rarely talk to each other, evaluating findings against the competence stack, being honest about what we know and what we don't, and building the outputs that make the synthesis actionable.

3.1 WHAT WE INTEND TO BUILD

The research program is not an end in itself. It serves two concrete output streams:

1. Curricula that provide a floor of capability. The highest-priority capabilities — self-regulation, honest self-assessment, the ability to learn new things independently, practical competence, relational capability — should be developed in every person, regardless of the educational setting. This is a capability floor, not a ceiling. The curriculum is organized around dimensions of adult competence (not traditional subjects), designed for the full lifespan (not just K–16), and grounded in the evidence about what actually produces durable capability.

2. Tools that embed learning science into the infrastructure. Spaced repetition foundations that can be woven into AI agents, so that the spacing and retrieval practice effects — among the most robust findings in all of psychology — are structurally delivered rather than left to individual discipline. Learning guides that translate the lab's findings into actionable advice for a person who wants to learn something right now. Assessment tools that separate the formative function (informing learning) from the summative function (evaluating for accountability).

4

THE COMPETENCE STACK

Everything in this document is evaluated against a five-layer model of what it means for a person to be genuinely competent. The stack is the lab's primary intellectual contribution — not as a novel theory, but as a synthesis of converging evidence from expertise research (Ericsson, Dreyfus), naturalistic decision-making (Klein), metacognition research (Flavell, Dunning & Kruger), and organizational learning (Edmondson, Argyris).

Layer 1: Domain Knowledge. The relevant facts, history, precedents, and constraints of a field. This is the most visible layer and the one traditional education most reliably addresses. It is necessary but radically insufficient on its own.

Layer 2: Skill. Knowledge made operational through practice, including the tacit knowledge (Polanyi) that a practitioner may not be able to articulate. The gap between knowing what good performance looks like and being able to produce it. Skill develops through deliberate practice with feedback, not through exposure alone.

Layer 3: Judgment. The capacity to determine *which* knowledge and skills to deploy in a given situation. Judgment requires calibrated mental models, sensitivity to second-order effects, and the ability to distinguish signal from noise. It cannot be directly transmitted. It must be developed through exposure to varied, consequential, and ambiguous situations with structured feedback. Kahneman and Klein — who disagree about almost everything else — agreed on these two conditions.

Layer 4: Metacognition. Awareness of one's own cognitive processes: the ability to monitor one's own performance, recognize the boundaries of one's knowledge, and update beliefs in response to evidence. The Dunning-Kruger effect operates here: the skills required to produce good work are the same skills required to recognize good work, so deficits at this layer are self-concealing. Metacognitive training has moderate effect sizes and is not a luxury — it is a prerequisite for learners to accept effective instruction, because effective learning strategies feel harder than ineffective ones.

Layer 5: Character and Disposition. Intellectual honesty. Tolerance for uncertainty. Courage to deliver or receive bad news. Willingness to say "I don't know." The habit of engaging with reality rather than performing confidence. These are not fixed personality traits but epistemic dispositions that can be cultivated or degraded by environment.

The stack is hierarchical but not additive. A person with intact Layers 1–2 but failed Layers 3–5 can do enormous damage — and the damage is often invisible to evaluation systems that only test the lower layers.

The fifth diagnostic question — "are they allowed to tell the truth in that system?" — is the environmental multiplier. It is not merely additive. Environments that penalize honesty do not just suppress truth-telling; over time they degrade the capacity to perceive truth. Diane Vaughan documented this as "normalization of deviance" in her analysis of the Challenger disaster. Amy Edmondson (10,040 citations) established that psychological safety is a prerequisite for team learning. Argyris showed that most organizations systematically prevent the double-loop learning (questioning goals and assumptions, not just adjusting actions) that upper-layer competence requires.

Part II

THE METHOD

5 HOW WE DID THE RESEARCH

5.1 THE AGENT ARCHITECTURE

The lab uses Claude Code — Anthropic's AI assistant — as its primary research infrastructure. Each investigation is conducted by an AI agent operating within a structured mandate: a set of research questions, quality standards, access to academic search tools, and explicit instructions about intellectual honesty. The Principal Investigator sets direction, reviews output, connects findings across agents, and maintains the lab's normative commitments.

The research proceeded in two waves:

Wave 0 produced a single full-field survey. One agent mapped the entire landscape of pedagogy and learning science, covering ten major domains across 51 typeset pages.

Wave 1 dispatched ten agents, one per domain:

1. Cognitive Foundations
2. Motivation and Self-Regulation
3. Assessment and Feedback
4. Instructional Design
5. Educational Technology and AI
6. Alternative Education
7. Institutional Analysis
8. What Should Be Learned
9. Competence Formation
10. Training Science

Each agent produced a literature review, an annotated bibliography with confidence ratings, a gap analysis, and practical implications for curriculum design. The total output is roughly 478 pages across all agents. Every review has been typeset and published.

5.2 THE TOOLS WE BUILT

The Scholar CLI. A command-line tool wrapping the OpenAlex and Semantic Scholar APIs. Agents use it to search for papers, retrieve abstracts, check citation counts and field-weighted citation impact, verify retraction status, find semantic recommendations, and access open-access full texts. It also searches Open Library and Project Gutenberg for books, and the Internet Archive for older texts.

The Reading Guide Pipeline. A separate project that transforms books into calibrated reading guides. The pipeline processes EPUBs through section-level summarization, then generates reading

guides and agent briefs tailored to the PI's intellectual program. As of April 2026, the library contains 55 books — from Dewey and Nussbaum through Kapur and Clark — each with structured summaries. Agent briefs (roughly 2–3K words each, roughly 50x smaller than the full books) provide numbered claims with evidence quality ratings, methodological notes, limitations, and connections to other literature. These briefs are optimized for agent consumption and are the primary mechanism by which the lab's agents engage with book-length sources.

Learning Guides. Distillations of the lab's findings into actionable advice for a person who wants to learn something. The first learning guide — seven principles drawn from all ten L1 reviews — was produced in March 2026.

The CCA Experiment. The lab's first live application of its own findings: using Applied Pedagogy's learning science principles to study for the Anthropic Claude Certified Architect certification. The experiment applies the expertise-adaptive model, productive failure, retrieval practice, and structured self-assessment to real-world skill acquisition — and documents what works and what doesn't.

5.3 WHAT THIS METHODOLOGY IS AND IS NOT

What it does well. It covers enormous ground quickly. A single agent can survey a field that would take a human researcher months, identifying key papers, tracking citation networks, checking retraction status, and synthesizing across hundreds of sources. The breadth of the Wave 1 output — ten domains, each reviewed in depth — would be difficult for a single human researcher to produce in the same timeframe.

What it does honestly. Every review includes an explicit gap analysis. "We don't know" appears in every review, and it is always the most trustworthy part. Several popular education findings — growth mindset interventions ($d = 0.02$–0.05 after bias correction), grit as an independent construct, learning styles — were identified and flagged as having weak or contradicted evidence. The lab is more skeptical of its own output than a typical academic lab would be: overnight spot-checks, a formal verification framework with documented failure modes, and a three-level provenance standard for claims.

What it does not do. The agents largely did not read the documents they cited. This is the most important limitation and it must be stated clearly. Most claims in the Wave 1 reviews trace to the AI model's training knowledge, not to a source the agent opened and verified in the current session. The citations are real — the OpenAlex IDs check out, the authors and dates are correct — but the citation format does not distinguish "I read this on page 47" from "I'm fairly sure this paper says this based on my training data."

The lab has adopted a three-level provenance standard going forward: **Verified** (agent read the document, claim traced to specific pages, local copy exists), **Abstract-verified** (agent read the abstract and metadata via the scholar tool), and **Training-derived** (claim comes from the model's training knowledge). The vast majority of Wave 1 claims are training-derived. This does not mean they are wrong — the model's training knowledge is usually accurate — but it means they have not been independently verified against the original sources.

A hunch from training data with no identifiable source is still a perfectly valid thing to include and to make decisions with. The lab values the model's pattern recognition across enormous amounts of literature — that is one of its strengths. But a hunch must be labeled as a hunch. "This seems likely based on patterns in the literature, but I cannot point to a specific source" is an honest and useful statement. Suppressing hunches because they can't be cited would make the research worse, not better.

There is a scientific opportunity here. The comparison between L0-001 (no API access, training knowledge only) and L0-001v2 (the same task with live academic databases) constitutes a natural experiment in model reliability — same task, same model family, different access to sources. Systematic comparison of the two outputs can characterize where training knowledge is accurate, where it confabulates, and where it simply has gaps. This extends further: as we work through the L1 agent outputs and attempt to discover the provenance of every claim the agents made without citation, we will accumulate a substantial dataset on what the models actually know versus what they merely sound confident about. We expect this process to be illuminating — and we are honest about being both nervous and excited about what it will reveal.

Part III

THE FINDINGS

6 WHAT WE FOUND

The findings are organized by the competence stack, not by the domain boundaries the agents used. This reframing reflects a lesson learned during Wave 1: the most important findings live at the intersections of domains, not within them.

6.1 KNOWLEDGE AND SKILL (LAYERS 1–2)

The science of how humans acquire and retain knowledge is the most mature part of the field. The key findings are robust, well replicated, and actionable. They are also, for the most part, ignored by the institutions that teach.

Working memory is the bottleneck. Conscious processing handles roughly four chunks of novel information at a time (Cowan, 2001). Not seven, as the popular version of Miller's 1956 paper claims — four. This biological constraint is what makes learning effortful. Cognitive load theory, built on this foundation, has a 2019 retrospective with 1,740 citations and an FWCI of 106. It is not controversial in the research community. It is, however, routinely violated in practice: teachers present twelve new concepts in a fifty-minute period, textbooks embed decorative images that consume processing without aiding learning, and ed-tech companies design interfaces that maximize visual engagement at the expense of cognitive processing.

Retrieval practice is the strongest single learning strategy. Actively retrieving information from memory strengthens it far more than re-exposure. Effect sizes in authentic classrooms are roughly 0.5 (Yang, Luo, Chu & Geng, 2021) — confirmed across 272 comparisons (Adesope et al., 2017). Free-recall formats produce larger effects than recognition (multiple-choice) because they require more effortful processing. The optimal initial success rate is 50–80% — hard enough to be effortful, not so hard as to produce only errors. Feedback after retrieval is essential for error correction.

But here is the devastating meta-finding: students systematically avoid retrieval practice because effort feels like failure (Kirk-Johnson, Galla & Fraundorf, 2019). They gravitate instead toward rereading and highlighting — strategies that Dunlosky and colleagues (2013) rated as having the lowest utility of any study strategy evaluated. The reason is that desirable difficulties (Bjork, 1994) and undesirable difficulties feel identical subjectively (Lodge, Kennedy & Lockyer, 2018). Students cannot tell whether struggling means "this is working" or "this is a waste of time." This single finding has an enormous practical implication: effective learning strategies must be built into curriculum structure, not left to student choice. If you let students choose how to study, they will choose the method that produces the least learning, every time, because it feels the most productive.

Spacing and interleaving multiply the effect. Distributing practice across time (spacing) and mixing different problem types during practice (interleaving) produce durable learning gains. The heuristic for spacing intervals is 10–30% of the desired retention period. Interleaving feels harder and produces more errors during practice — but substantially better long-term retention and discrimination. These effects are robust across domains.

The worked example effect transitions to fading. Novices learn more from studying worked examples than from solving equivalent problems (Barbieri, Booth & Begolli, 2023 meta-analysis). As

competence increases, examples should fade — first to completion problems, then to independent practice. The 4C/ID model (van Merrienboer & Kirschner) extends this to complex skill domains, insisting on whole-task practice from the beginning with simplified but authentic problems.

What works for novices harms experts, and vice versa. The expertise reversal effect is among the most important findings for curriculum design. A 2025 meta-analysis (Tetzlaff, Simonsmeier, Peters & Brod) across 176 effect sizes: low-knowledge learners benefit from high assistance ($d = 0.505$) while high-knowledge learners benefit from low assistance ($d = -0.428$). The asymmetry is important: providing too much scaffolding is less harmful than providing too little. One-size-fits-all instruction is guaranteed suboptimal for most learners.

The direct instruction versus inquiry debate is a false binary. This may be the most consequential finding for anyone designing curricula. The debate has consumed decades and enormous political energy. Thirteen leading researchers — including figures from both sides — converged in 2023: "a combination of inquiry and direct instruction may often be the best approach." The outcome measure determines which approach "wins": direct instruction wins on immediate factual recall; inquiry wins on conceptual understanding, transfer, and long-term retention. The resolution is not compromise but sequence — an expertise-adaptive model: explicit instruction for novices, faded scaffolding for beginners, productive failure for intermediates, guided inquiry for advanced learners.

Deliberate practice develops skill, but explains less variance than its advocates claimed. Ericsson's framework (8,634 citations) established that expert performance requires structured practice with feedback. But Macnamara and colleagues' 2014 meta-analysis found deliberate practice accounts for only 26% of variance in games, 21% in music, and less than 1% in professions. The "10,000 hour rule" oversells it.

Spaced repetition software is the highest-confidence technology application — computationally implementing two of the most robust findings in psychology. It addresses the metacognitive error structurally: it imposes spaced retrieval practice regardless of learner preferences. But it is a precision tool for declarative knowledge only. It cannot develop skill, judgment, metacognition, or character.

6.2 JUDGMENT (LAYER 3)

This is where the evidence thins. Cognitive science has "almost nothing to say" about judgment directly, as our cognitive foundations review acknowledged.

The Kahneman-Klein conditions specify when judgment develops: high environmental validity (stable, learnable regularities) plus adequate feedback opportunity (clear, timely feedback on judgment outcomes). Without feedback, experience does not correct errors; it entrenches them. Most classrooms provide neither the validity nor the feedback that judgment requires.

Training contexts develop judgment far better than schools. Military simulation, aviation CRM, and medical scenario-based training all provide what the Kahneman-Klein conditions specify. The structural explanation: when the organization bears the cost of failure (crashed aircraft, dead patients), the incentive structure forces diagnostic rather than evaluative assessment, flexible time-to-mastery, fast feedback, and tolerance for errors during practice. Schools have exactly the opposite structure.

Productive failure is the most promising bridge. Manu Kapur's paradigm reliably produces better conceptual understanding and transfer than direct instruction alone — effects up to two academic years for transfer outcomes. The mechanism is not that struggle alone produces learning;

it is that struggle followed by instruction produces deeper learning than instruction alone. But productive failure is primarily tested in mathematics.

After-action reviews produce large effects ($d \approx 0.67$ in Tannenbaum & Cerasoli's 2012 meta-analysis) with a simple four-question structure. The cultural norms matter as much as the structure: rank-neutral, honest self-assessment expected, focus on processes not blame. This connects directly to the post-mortem traditions at Amazon and Google, and to NASA's SP-287 analysis of what made Apollo succeed — convergent evolution across very different institutions arriving at the same practices because they work.

6.3 METACOGNITION (LAYER 4)

Metacognitive judgments are systematically wrong. Students in active learning sections at Harvard learned more but reported lower satisfaction and lower perceived learning than students in passive lecture sections (Deslauriers et al., 2019). Effective learning feels harder than ineffective learning. Student satisfaction cannot be trusted as a quality metric.

Self-regulation is the meta-capability. The Dunedin study (Moffitt et al., 2011): childhood self-control predicted adult outcomes across health, finance, and criminal behavior in a continuous gradient, even after controlling for IQ and social class, from birth to age 32. The Perry Preschool Study: effects on IQ faded within years but effects on self-regulation persisted and produced better outcomes four decades later. Heckman's economics of skill formation: "skills beget skills" — early investments create platforms for later development. Self-regulation can be taught through direct instruction but does not transfer automatically across domains.

Feedback literacy must be explicitly taught. The capacity to use feedback productively — appreciating it as information, managing the emotional response, converting it into changed behavior — is not something students arrive with (Carless & Boud, 2018, 1,800 citations).

6.4 CHARACTER AND DISPOSITION (LAYER 5)

This is where the evidence is thinnest and where Applied Pedagogy's distinctive contribution is intended to lie.

Intellectual humility can be measured but cannot yet be directly trained. The honest answer is that we do not know how to teach intellectual honesty or tolerance for uncertainty through direct instruction. The evidence points toward environmental design rather than curriculum.

Environment is the primary mechanism for character development. This is the strongest cross-cutting finding in the entire lab — five independent agents, studying different literatures, arrived at the same conclusion. Environments that reward honesty, treat error as information, and tolerate uncertainty develop epistemic character. Environments that penalize honesty, reward performance of confidence, and treat error as failure degrade it. The conventional school environment — grading systems that punish honest self-assessment, hidden curricula that teach compliance over inquiry, institutions that implicitly value confident performance over honest engagement — works against character development at every turn.

6.5 THE ENVIRONMENTAL MULTIPLIER

Self-determination theory (Ryan & Deci, 2000; 27,000+ citations): environments that thwart autonomy, competence, and relatedness produce not just reduced motivation but active defiance, anxiety, and helplessness. The assessment literature: grades negate the benefit of feedback (Butler,

1988); high-stakes testing creates controlling environments incompatible with deep learning. The institutional analysis: the "grammar of schooling" (Tyack & Cuban, 1995) has survived every reform wave for a century. The training science literature: psychological safety is literally training infrastructure. The alternative education review: Montessori's absence of extrinsic rewards aligns with SDT's prescriptions and is associated with better executive function and social-cognitive outcomes (Lillard & Else-Quest, 2006, *Science*).

Environmental design precedes curriculum design. Before specifying content, methods, or assessments, design the environment: its reward structures, error tolerance, authority relationships, and what happens when someone says "I don't know."

6.6 WHAT TECHNOLOGY GETS WRONG

This is where the lab's findings become urgent for the current moment.

More technology can mean less learning. Immersive VR produced more presence and engagement but significantly less learning (Makransky, Terkildsen & Mayer, 2019). The mechanism: VR consumed cognitive resources that masqueraded as engagement. Gamification effects are small ($d \approx 0.36$) with significant heterogeneity, and points-badges-leaderboards function as extrinsic rewards that risk undermining intrinsic motivation. Ed-tech companies optimize for engagement metrics — time-on-task, completion rates, session frequency — that do not measure learning. The market systematically selects for pedagogically inferior products.

Over one-third of feedback interventions make things worse. Kluger and DeNisi (1996) found across 607 effect sizes that 38% of feedback interventions *decreased* performance. Feedback becomes harmful as attention moves from the task level to the self level. This is not a minor finding. It means that most of what passes for feedback in education — "good job," "B+," praise directed at the person rather than the work — is at best useless and at worst actively counterproductive.

The LLM evidence base is essentially zero. As of early 2025, there are no randomized controlled trials examining the effect of LLM-based tutoring on learning outcomes. Adoption has outpaced evidence by orders of magnitude. The underlying cognitive science makes directional predictions, and they are not encouraging: the testing effect predicts easy access to answers reduces learning; productive failure research predicts immediate help reduces conceptual understanding; the generation effect predicts that receiving perfect explanations is worse than producing imperfect ones. A student who produces polished AI-assisted output may have the impression of competence where competence does not exist.[1]

Intelligent tutoring systems get roughly halfway to Bloom's goal. ITS produce $d \approx 0.66$ compared to conventional instruction (Kulik & Fletcher, 2016). But VanLehn's (2011) key distinction matters: step-level systems (that interact during problem-solving) produce $d \approx 0.76$, comparable to human tutoring, while problem-level systems (that evaluate after a complete answer) produce effects indistinguishable from zero. The inner loop — moment-by-moment interaction with learner reasoning — is where learning happens.

6.7 THE CROSS-CUTTING THEMES

The assessment-motivation tension. Assessment is simultaneously the most powerful lever for learning and the most common mechanism for undermining motivation. Extrinsic rewards (grades,

[1] The irony of this observation appearing in a document 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.

competitive rankings) reliably undermine intrinsic motivation — a meta-analysis of 128 experiments (Deci, Koestner & Ryan, 1999). Grades negate the benefit of feedback comments (Butler, 1988): students receiving comments-only showed the highest subsequent interest and performance; grades-only showed the lowest; comments-plus-grades performed the same as grades-only. The grade overwhelms the feedback. The resolution: separate the formative function of assessment from the summative function. Use retrieval practice for learning, not grading.

Motivational decline across schooling. Intrinsic motivation declines systematically from childhood through adolescence across all subjects. The stage-environment fit hypothesis: middle schools become more controlling precisely when adolescents need more autonomy. This is one of the most important and depressing findings in the field, and its causal mechanisms remain underresearched.

The training-education bridge. CRM is arguably the most successful training intervention in history — a 20-fold reduction in commercial aviation accident rates. Education researchers barely notice because training science lives in different journals. The structural insight: when the organization bears the cost of failure, everything about learning design changes.

The knowledge-skills false dichotomy. You cannot think critically without domain knowledge, and domain knowledge without thinking capacity is inert (Willingham, 2009). Transfer is domain-specific and limited. "Critical thinking" as a domain-independent skill is a questionable construct.

Errors are information, not failure. Productive failure shows initial errors produce deeper subsequent learning. Error management training produces superior transfer. Error management cultures produce better organizational performance. Yet the hidden curriculum of schools teaches that errors are evidence of inadequacy.

Alternative education evidence is structurally compromised. The alternative education review produced the lab's most uncomfortable honest finding: we are essentially flying blind. Selection bias in homeschooling research is arguably irremediable — families who choose alternatives differ on unmeasured characteristics that are almost certainly the dominant factors. Montessori is the only alternative model with genuine evidence (Lillard & Else-Quest, 2006, lottery-based study in *Science*), and its specific principles — absence of extrinsic rewards, extended uninterrupted work periods, self-correcting materials — align with both SDT and cognitive science. The structured-versus-unstructured distinction is the critical variable (Martin-Chang, Gould & Meuse, 2011): structured homeschooling scored significantly higher than matched conventional peers; unstructured homeschooling scored significantly lower. The alternative education community and the conventional education establishment both have captured research; the honest broker's position requires skepticism toward all camps.

The grammar of schooling as coordinated equilibrium. Every country in the world converges on the same basic school architecture — age-graded classrooms, subject silos, exam-based credentialing, bell schedules — not because it is optimal but because deviating imposes coordination costs that no individual actor can absorb. This is a Nash equilibrium, not a designed optimum. Nobody sat down in 1850 and concluded that age-graded batch processing was the ideal way to teach children — it emerged, spread, and locked in. The result is that the global school system optimizes for compatibility with itself, not for learning. COVID was the test: institutions had every reason to adapt, enormous pressure to experiment, and they reverted to baseline the moment they could.

7 WHAT WE ARE CONFIDENT ABOUT AND WHAT WE ARE NOT

7.1 HIGH CONFIDENCE

- Retrieval practice, spacing, and interleaving should be structurally incorporated into all curricula.
- Extrinsic rewards reliably undermine intrinsic motivation. The standard grading apparatus is actively demotivating.
- The expertise reversal effect means instruction must adapt to learner expertise level.
- Growth mindset interventions are negligible ($d = 0.02$–0.05 after bias correction). The intervention paradigm is dead.
- Self-regulation is the most robustly predictive meta-capability and can be taught through direct instruction.
- The institutional environment is a first-order determinant of learning outcomes at Layers 3–5.
- The science-practice gap in education is structural, not informational.
- Multiple philosophical traditions converge on the same core dimensions of adult competence.

7.2 MEDIUM CONFIDENCE

- Productive failure reliably outperforms direct instruction for conceptual understanding in mathematics. Extension to other domains is plausible but undemonstrated.
- Formative assessment improves learning, but the effect size is debated ($d = 0.20$ to $d = 0.70$).
- After-action reviews produce large effects, but nearly all evidence comes from training contexts.
- The motivational decline across schooling is primarily institutional. The stage-environment fit hypothesis is compelling but causal mechanisms are underresearched.

7.3 LOW CONFIDENCE OR GENUINELY UNKNOWN

- Whether any approach produces reliable far transfer.
- Whether LLM-based tutoring helps or harms learning. There are essentially no rigorous outcome studies.
- How to teach effectively in ill-structured domains (writing, ethics, design).

- Whether intellectual humility can be directly trained.
- Whether any of these interventions compound, plateau, or fade over years.
- Whether these findings generalize beyond Western, educated, industrialized populations. Seven of ten agents independently flagged this gap.
- How to assess Layers 3–5 with institutional-grade reliability.

Part IV

ASSESSMENT AND FUTURE

8

THE LIMITATIONS

These limitations are load-bearing. A research program that does not know its weaknesses cannot be trusted to know its strengths.

Source provenance. The vast majority of Wave 1 claims trace to the AI model's training knowledge rather than to sources the agents actually read. The citations are real, but the agents did not verify most claims against the original text. The lab's provenance standard distinguishes verified, abstract-verified, and training-derived claims. The long-term goal is full verification with page-level citations.

Source discovery bias. Reviews are biased toward peer-reviewed English-language publications. Books and monographs are underrepresented. Practitioner literature, grey literature, and non-English sources are largely absent.

Conventional domain boundaries. The agents were organized along traditional academic lines. The most valuable findings live at the intersections. A different organizational scheme — by question rather than by domain — might have produced more integrated findings.

No adversarial review. No agent was tasked with finding where the lab is wrong. Given the model's tendency toward coherent narratives, this matters.

WEIRD population bias. Nearly the entire evidence base comes from Western, educated, industrialized, rich, and democratic populations.

The long-term gap. Most studies measure weeks. A curriculum is a multi-year undertaking. Whether these interventions compound or fade over years is essentially unknown.

9

THE OPEN INVESTIGATIONS

The lab has generated a substantial agenda of questions that emerged during the reading of the Wave 1 reviews. These are not abstract academic questions — they are the questions the PI needs answered to build what Applied Pedagogy intends to build.

What do teacher colleges actually teach? The lab has produced ten domain reviews. Teacher preparation programs have their own version of this knowledge. Where does the lab agree with, overlap with, disagree with, and diverge entirely from what teacher candidates are taught? The comparison tells us where Applied Pedagogy adds value and where it is redundant. Ball State's Teachers College is five minutes away.

What are the consequences if we screw up? If Applied Pedagogy is going to recommend that people deviate from conventional education, it must be honest about the risks. What does the evidence actually say about the downside of educational experimentation? What is actually hard to recover from versus what is easily remediated? The competence stack predicts that upper-layer deficits (judgment, metacognition, character) may be harder to remediate than lower-layer ones (knowledge, skill) — a kid who can't do algebra but has excellent self-direction can learn algebra later, but a kid with algebra and no self-direction may never learn anything voluntarily again.

The wasted COVID experiment. COVID was the biggest education disruption in a century. Every school was forced to stop operating normally. The institutions reverted to baseline the moment they could — confirming the grammar-of-schooling prediction. Is there still salvageable data? Are there families who discovered homeschooling worked and kept doing it? Is anyone asking the right questions about the COVID cohort, or is the entire research community framing it as "learning loss to recover from" rather than "natural experiment to learn from"?

Predictive processing as a unifying theory. The PI has been reading Andy Clark and related authors. PP — the brain as a prediction-error minimization machine — keeps connecting to everything the lab found. Productive failure works because wrong predictions generate error signals. Retrieval practice works because it forces predictions from memory. The expertise reversal effect is what happens when your model is already good enough that easy examples generate no signal. PP could be the missing "why" underneath all the "what works" findings.

How do learners update wrong knowledge? The lab's own operational problem — what if an agent mischaracterized a source? — is an instance of a general pedagogical problem. The conceptual change literature and interference theory suggest that simply telling someone a fact was wrong is insufficient. You need to actively practice retrieving the corrected version at spaced intervals. Any curriculum system should have a mechanism for propagating corrections — and that mechanism should not stop when the student graduates. A person should be able to discover, at forty, that something they learned at twenty was wrong, and have a way to actively replace it.

Video game tutorials as instructional design. Game tutorials represent a sophisticated, commercially tested tradition of instructional design that education researchers ignore. They teach complex systems without manuals, without grades, and without the player feeling "taught." They are criterion-referenced by design. They maintain motivation through the flow channel. The design principles may translate.

Post-mortems, blameless culture, and high-competence organizations. The Amazon COE process, Google's blameless post-mortems, and NASA SP-287 ("What Made Apollo a Success") are positive exemplars of the competence stack at institutional scale. They are convergent evolution —

different organizations independently arriving at the same practices because they work: blameless analysis, systematic decomposition, causal reasoning, institutional memory.

10 WHAT COMES NEXT

10.1 PHASE 1: THIS DOCUMENT

The retrospective you are reading now. It synthesizes the Wave 1 output, states the limitations honestly, and provides a collaborator-ready entry point to the project.

10.2 PHASE 2: REFINED REVIEWS

The lab will go back through each domain with improved methods — deeper source reading, book engagement at summary depth via the agent brief pipeline, explicit counter-evidence searches, non-academic sources where academic evidence is thin. Seven agents instead of ten, merging deeply intertwined domains: cognition with instructional design, motivation with assessment, competence formation with training science, curriculum with philosophy.

Each agent produces a complete refined review (v2) alongside a changelog documenting what changed. The original reviews are preserved. The full intellectual trajectory — v1 reviews, v2 reviews, changelogs, and this retrospective — will be published at lab.appliedpedagogy.com. We do not hide the first attempt. The trajectory is the evidence that the methodology improves.

10.3 PHASE 3: COMPARATIVE ANALYSIS

A single agent reads all v1 and v2 reviews and produces a structured assessment: where did the foundation hold up, where did it crack, what is genuinely new, what is still thin? This phase includes the adversarial function that Wave 1 lacked — looking for contradictions, citation chains posing as consensus, and systematic biases.

10.4 PHASE 4: WAVE 3

Genuinely new questions rather than refinement. The structural commitments: organize by question rather than by domain, front-load the normative framework, include adversarial agents, and allow investigations that pull from whatever domains they need.

11

WHAT IT ALL MEANS

The science of learning is real and actionable. We know how to help people acquire knowledge and develop skill — the lower layers of the competence stack — with reasonable confidence. Retrieval practice, spacing, interleaving, worked examples for novices, faded scaffolding, formative assessment, and autonomy-supportive instruction are all well-supported.

But knowledge and skill are not enough. The layers that matter most — judgment, metacognition, and epistemic character — are the layers that existing educational systems systematically fail to develop. They fail not because nobody cares but because the institutional structures that make schools function are the same structures that prevent the development of the upper layers. The grammar of schooling is optimized for Layers 1–2 and is structurally hostile to Layers 3–5.

Training contexts — military, aviation, medicine — have solved parts of this problem, because they must. When the organization bears the cost of failure, the incentive structure forces attention to the full stack. Education has barely noticed.

The environmental dimension is the linchpin. Environments that reward honesty, treat error as information, and tolerate uncertainty develop competence. Environments that reward performance of confidence, punish error, and demand certainty degrade it. This is not a secondary consideration. It is the single most important design variable.

Education is not something that happens to you for sixteen years and then stops. It is the development of capabilities to move through life — capabilities that change across the lifespan, both in what is required and in what a person becomes capable of acquiring. An educational approach that cannot produce a person who continues to learn after the instruction stops has failed at the thing that matters most.

Applied Pedagogy's contribution is the synthesis: bringing these findings together across domains that rarely talk to each other, evaluating them against a coherent framework of what competence actually means, being honest about what we know and what we don't, and building tools and curricula that make the synthesis actionable. The work is incomplete. The limitations are real. But the alternative to doing this work imperfectly is not doing it at all — and no existing institution is positioned to do it. The science exists. The synthesis exists. The question is whether anyone will act on it.

This document was produced by the Applied Pedagogy Research Lab in April 2026. Applied Pedagogy will likely be organized as an Indiana Public Benefit Corporation. The full reviews, annotated bibliographies, and gap analyses are available at LAB.APPLIEDPEDAGOGY.COM. *The lab's methodology, quality standards, and normative commitments are documented in* LAB.MD *and* COMPETENCE-TARGET.MD *at the project root.*