BUILDING THE WHOLE PERSON

How Competence Is Formed, Maintained, and Degraded

Applied Pedagogy Research Lab

Guido Bartolucci, Principal Investigator

guido@appliedpedagogy.com

LAB.APPLIEDPEDAGOGY.COM

L1-009 · March 2026

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

CONTENTS

Part I

FOUNDATIONS

1

THE PROBLEM OF THE UPPER LAYERS

Educational systems are reasonably good at transmitting knowledge and building skills. Retrieval practice, spaced repetition, deliberate practice, worked examples — the cognitive science findings catalogued in the Lo survey and subsequent L1 investigations constitute a genuine science of instruction for the lower layers of the competence stack. A well-designed curriculum, informed by these findings, can reliably produce learners who know relevant facts and can perform relevant procedures. This is not a trivial achievement. It is the product of decades of careful research, and it works.

But knowledge and skill are not competence. They are necessary components — the foundation of the stack — but they are radically insufficient on their own. The surgeon who knows anatomy and can execute procedures but cannot judge when surgery is the wrong option is not competent. The engineer who can run calculations but cannot sense when a design is heading toward failure is not competent. The teacher who knows pedagogy but cannot read a classroom is not competent. What separates the competent from the merely credentialed is the capacity for judgment, self-monitoring, and epistemic honesty — layers 3, 4, and 5 of Applied Pedagogy's competence stack.

These upper layers are where educational systems fail most conspicuously, and they are where the evidence base is thinnest. This is not a coincidence. Layers 1 and 2 — domain knowledge and skill — are amenable to the kinds of standardized, scalable instruction that institutions are designed to deliver. They can be taught through lectures, textbooks, and practice problems. They can be assessed through tests and demonstrations. They have clear, measurable outcomes. The upper layers resist all of this. Judgment cannot be transmitted through instruction alone — it must be developed through exposure to varied, consequential, and ambiguous situations. Metacognition requires the learner to monitor a cognitive process that is, by definition, happening below the level of conscious awareness most of the time. Character and disposition are shaped by environment at least as much as by explicit training, which means that the institutional context in which education occurs is not merely a backdrop to learning but a first-order determinant of outcomes.

This investigation examines the empirical evidence for how each layer of the competence stack is developed, maintained, and degraded. It proceeds from the well-established (layers 1–2) through the moderately researched (layers 3–4) to the thinly evidenced (layer 5), and it treats the environmental dimension — the structural and institutional conditions that promote or undermine competence — as a first-order question rather than an afterthought. The investigation is grounded in COMPETENCE-TARGET.md, which defines the five-layer stack and the diagnostic questions that identify where competence breaks down. The central finding can be stated in advance: we know a great deal about building knowledge and skill, a moderate amount about developing judgment and metacognition, and very little about cultivating epistemic character — but the evidence we do have suggests that the environmental dimension, particularly the presence or absence of psychological safety and error-tolerance, may be the single most important factor in whether upper-layer competence develops or degrades.

2

LAYERS 1–2: THE ESTABLISHED SCIENCE OF KNOWLEDGE AND SKILL

2.1 THE KNOWLEDGE BASE

The science of how people acquire, store, and retrieve knowledge is the strongest area in all of educational research. Four findings stand on particularly firm empirical ground.

Retrieval practice — the act of recalling information from memory rather than simply re-reading it — produces substantially better long-term retention than passive review. Karpicke and Roediger's (2008) demonstration that free recall tests outperformed repeated study even without feedback has been replicated hundreds of times across age groups, domains, and settings. The effect is robust, large (typically d = 0.5–0.7), and practically useful. As the L1-003 assessment investigation documented, this finding undergirds the "testing effect" and provides the strongest evidence-based case for frequent, low-stakes quizzing in educational contexts.

Spaced repetition — distributing practice over time rather than massing it — produces more durable learning. The spacing effect has been known since Ebbinghaus (1885) and has been confirmed by more than a century of subsequent research. Cepeda et al. (2006) conducted a comprehensive meta-analysis confirming that distributed practice consistently outperforms massed practice for long-term retention, with optimal spacing intervals that increase as the target retention interval lengthens.

Cognitive load theory provides the theoretical framework for understanding why instruction fails when it does. Working memory is severely limited — Cowan (2001) estimated its capacity at approximately four items — and instruction that exceeds this capacity produces learning failure regardless of content quality. The practical implications (reduce extraneous load, manage intrinsic load through scaffolding, optimize germane load through meaningful processing) are well-supported, though primarily in well-structured domains (Sweller, Ayres & Kalyuga, 2011).

Deliberate practice — structured, effortful practice with feedback, targeting specific weaknesses — is the primary mechanism through which skill develops. Ericsson, Krampe, and Tesch-Römer (1993) identified it as the key factor distinguishing elite performers from competent practitioners in music, finding that expert violinists had accumulated approximately 10,000 hours of solitary practice by age twenty. This finding, with 8,634 citations and a field-weighted citation impact of 35.0, became one of the most influential in all of psychology — and one of the most oversimplified.

2.2 THE DELIBERATE PRACTICE DEBATE

The popularization of the "10,000 hour rule" — largely through Malcolm Gladwell's *Outliers* (2008) — distorted Ericsson's actual findings in two important ways. First, Ericsson never claimed that 10,000 hours was a fixed threshold; it was an average for a specific sample of violinists, and the variance was substantial. Second, and more importantly, Ericsson's framework emphasized the *quality* of practice — its deliberate, structured, feedback-rich nature — not merely its quantity. Ten thousand hours of unfocused practice does not produce expertise.

Macnamara, Hambrick, and Oswald (2014) conducted a meta-analysis that challenged the scope of deliberate practice's explanatory power. They found that deliberate practice accounted for only

26% of variance in performance in games, 21% in music, 18% in sports, 4% in education, and less than 1% in professions. Ericsson and Harwell (2019) responded vigorously, arguing that the meta-analysis used overly broad definitions of "practice" that included activities Ericsson would not classify as genuinely deliberate. The debate remains active, but the consensus position is that deliberate practice is a necessary but not sufficient condition for developing expertise, and that its explanatory power varies substantially across domains — strongest in well-structured domains with clear feedback (chess, music) and weakest in ill-structured professional domains where the criteria for excellence are ambiguous.

This finding matters for the competence stack because it suggests that the mechanisms that build skill (layers 1–2) may have limited applicability to the upper layers. Deliberate practice works best when there are clear standards of performance, immediate feedback on errors, and well-defined sequences of increasingly difficult challenges. Judgment, metacognition, and character do not obviously meet these conditions.

2.3 THE SKILL-TO-JUDGMENT TRANSITION

The most important question for a competence stack is not how layers 1 and 2 work — that is reasonably well understood — but how competence develops *beyond* skill into the upper layers. Two frameworks address this transition directly.

The Dreyfus model of skill acquisition (Dreyfus & Dreyfus, 1986; S. Dreyfus, 2004) proposes five stages of development from novice to expert: novice, advanced beginner, competent, proficient, and expert. The critical transition occurs between the "competent" and "proficient" stages. At the competent stage, the practitioner follows rules and procedures deliberately and analytically. At the proficient stage, the practitioner begins to perceive situations holistically — to see patterns, recognize what is salient, and respond intuitively rather than analytically. At the expert stage, this intuitive perception becomes so refined that the practitioner "just sees" what needs to be done without conscious deliberation.

The Dreyfus model is a phenomenological description rather than an empirically tested causal theory, and this is both its strength and its limitation. Its strength is that it captures something practitioners across domains recognize as true: the transition from rule-following to intuitive perception is real, and it feels qualitatively different. Its limitation is that it does not specify the mechanisms by which this transition occurs, what environmental conditions support or impede it, or how it might be accelerated through instructional design. Stuart Dreyfus (2004), in a retrospective account with 1,205 citations, acknowledged that the model describes what happens but not how to make it happen.

Benner (2004) applied the Dreyfus model to nursing and provided the most detailed documentation of how the stages manifest in professional practice. Her descriptions of how nurses transition from applying rules ("if the patient's blood pressure drops below X, call the doctor") to perceiving clinical situations holistically ("something is wrong with this patient") have been enormously influential in healthcare education. Benner's work suggests that the transition requires two things: extensive experience with varied cases, and a practice environment that allows the practitioner to attend to the whole situation rather than reducing it to discrete variables. These conditions point toward judgment — layer 3 — as an emergent property of pattern recognition developed through experience, not a skill that can be directly taught.

The expert-novice paradigm in cognitive psychology provides the empirical complement to the Dreyfus model. Chi, Feltovich, and Glaser (1981) demonstrated that experts and novices categorize physics problems in fundamentally different ways: novices group problems by surface

features (problems involving inclined planes, problems involving springs), while experts group them by deep structural principles (problems involving conservation of energy, problems involving Newton's second law). This finding has been replicated across domains from medical diagnosis to chess to programming. It reveals that expertise is not simply "knowing more" but having qualitatively different mental representations — representations organized around deep structure rather than surface features.

The implication for the competence stack is profound. The transition from skill (layer 2) to judgment (layer 3) is not a matter of accumulating more knowledge or practicing more procedures. It requires a reorganization of mental representations — a shift from surface-level to deep-structural understanding that enables pattern recognition in novel situations. This reorganization cannot be directly transmitted through instruction. It must be developed through extensive, varied experience in which the learner confronts problems that require deep-structural analysis. Chi's work suggests that the process can be supported by instructional techniques that make deep structure visible — such as asking learners to compare and contrast cases, explain underlying principles, or categorize problems by structural features rather than surface features — but the reorganization itself must happen within the learner's own cognitive system.

Part II

THE UPPER LAYERS

3

LAYER 3: THE DEVELOPMENT OF JUDGMENT

Judgment — the capacity to determine which knowledge and skills to deploy in a given situation, to distinguish signal from noise, and to anticipate second-order effects — is the layer that separates the competent from the merely credentialed. It is also the layer where the educational research base begins to thin substantially. The most relevant literatures are naturalistic decision-making, the conditions for intuitive expertise, and the training of clinical judgment in professional domains.

3.1 NATURALISTIC DECISION-MAKING: HOW EXPERTS ACTUALLY DECIDE

Gary Klein's research program on naturalistic decision-making (NDM) emerged from a simple but powerful observation: the classical model of rational decision-making — listing options, weighing pros and cons, selecting the optimal choice — did not describe how experienced practitioners actually made decisions in complex, time-pressured, high-stakes situations. Klein studied fireground commanders, intensive care unit nurses, military commanders, and other expert practitioners in their natural environments and found that they rarely compared options at all. Instead, they recognized the situation as a type, generated a plausible course of action based on that recognition, mentally simulated the action to check for problems, and then either executed it or modified it. Klein called this the Recognition-Primed Decision (RPD) model (Klein, 1998).

The RPD model has three key implications for competence formation:

First, **expert judgment is fundamentally pattern-based**. Experts do not reason their way to decisions analytically; they recognize patterns in the current situation that map onto patterns stored in memory from previous experience. This means that judgment development requires extensive exposure to varied situations — not in the abstract but in contexts that allow the development of a rich repertoire of patterns. Klein estimated that experienced fireground commanders had a library of several hundred prototypical situation-action patterns that they could recognize and deploy within seconds.

Second, **mental simulation is the core mechanism of judgment quality**. What distinguishes good judgment from bad judgment is not the speed of pattern recognition but the quality of the mental simulation that follows it. After recognizing a situation and generating an initial course of action, the expert mentally "runs the movie forward" — imagining how the action will play out, looking for problems, checking whether the outcomes are acceptable. This simulation capacity depends on having accurate mental models of how the domain works — models that include causal relationships, temporal dynamics, and potential failure modes. Experts with good judgment have not just more experience but more accurately calibrated mental models.

Third, **judgment operates largely below the level of conscious awareness**, which means it cannot be directly transmitted through instruction. Klein found that experts often could not articulate why they made the decisions they made. When asked, they typically reconstructed rational-sounding explanations after the fact — explanations that did not accurately describe the actual cognitive process. This has been confirmed by decades of research on implicit knowledge and automaticity: expertise involves the development of fast, automatic recognition processes that bypass deliberate analysis (Polanyi, 1966; Kahneman, 2011).

3.2 CONDITIONS FOR INTUITIVE EXPERTISE: WHEN JUDGMENT IS RELIABLE

The most important contribution to the judgment question came from an unlikely collaboration between Daniel Kahneman and Gary Klein — two researchers with fundamentally opposed views on the reliability of expert intuition. Kahneman, the apostle of cognitive biases, had spent his career documenting the systematic errors in human judgment. Klein, the champion of NDM, had spent his career documenting the remarkable accuracy of expert intuition. Their 2009 paper, "Conditions for Intuitive Expertise: A Failure to Disagree" (2,324 citations, FWCI 33.44), represents one of the most important integrations in the judgment literature.

Kahneman and Klein agreed on two conditions that must be present for expert intuition to develop reliably:

High validity of the environment. The environment must contain stable, learnable regularities — patterns that repeat with sufficient consistency to be internalized. Chess is a high-validity environment: board positions have stable strategic implications that can be learned through experience. Stock markets are low-validity environments: the patterns are too noisy and unstable for reliable intuition to develop (which is why expert stock pickers consistently fail to beat indices). Medicine occupies an intermediate position: some diagnostic situations are highly patterned (a classic presentation of appendicitis), while others involve too many interacting variables for reliable pattern recognition (predicting which patients will respond to which chemotherapy regimes).

Adequate opportunity for learning. The practitioner must receive clear, timely feedback on the outcomes of their judgments. A physician who orders tests and sees the results the next day is in a good learning environment. A radiologist who reads films and never learns the surgical findings is in a poor one. A psychiatrist whose patients drop out of treatment and are never followed up is in a terrible one. Without feedback, experience does not correct errors — it entrenches them. Extensive experience in a low-feedback environment produces not expertise but "overlearned" errors with high confidence — a particularly dangerous combination.

The Kahneman-Klein framework implies that **judgment is not universally trainable**. In high-validity, high-feedback environments, judgment develops naturally from accumulated experience — and can potentially be accelerated through instructional designs that provide concentrated, varied exposure. In low-validity or low-feedback environments, judgment may not develop at all, no matter how extensive the experience, because the conditions for reliable pattern learning are absent.

This finding has a direct implication for the competence stack. Layer 3 (judgment) is not a generic capacity that develops uniformly across all domains. It is domain-specific and environment-dependent. The question "can judgment be taught?" does not have a single answer. In high-validity, high-feedback domains, the answer is "judgment develops through structured experience, and instructional design can accelerate this development." In low-validity or low-feedback domains, the answer is "judgment may not be achievable through experience alone, and the appropriate response is not better training but better decision-making structures" — checklists, decision aids, peer review, structured protocols that substitute institutional processes for individual judgment.

3.3 CASE-BASED REASONING AND SIMULATION: ACCELERATING JUDGMENT DEVELOPMENT

Several professions have developed instructional approaches specifically aimed at developing judgment, with varying degrees of empirical support.

Medical case-based reasoning. Medical education has the longest tradition of deliberately developing clinical judgment. The case method — presenting students with complex clinical scenarios that require diagnosis, differential reasoning, and treatment decisions — has been a cornerstone of medical education since the early twentieth century. The evidence suggests that case-based learning develops diagnostic reasoning and clinical judgment more effectively than lecture-based instruction, particularly when cases are varied (to prevent over-reliance on a narrow pattern library), discussed collaboratively (to expose students to reasoning processes different from their own), and followed by feedback on diagnostic accuracy (to calibrate mental models). Problem-based learning (PBL), which takes the case method further by making the case the primary vehicle for learning content as well as judgment, has shown mixed results in meta-analyses — generally equivalent to traditional instruction for knowledge acquisition but superior for developing clinical reasoning and diagnostic skills (Strobel & van Barneveld, 2009).

Military after-action reviews. The U.S. Army's after-action review (AAR) process, developed at the National Training Center in the 1980s, is one of the most systematic approaches to developing judgment through structured reflection on experience. After each training exercise, units conduct a detailed review of what happened, what was supposed to happen, what went right, and what went wrong — without regard to rank or position. The AAR process embeds several principles that the judgment literature suggests are important: it provides immediate, specific feedback tied to concrete situations; it makes decision-making processes explicit and subject to scrutiny; it creates a psychologically safe context for error examination; and it forces participants to articulate their mental models and compare them to reality. While controlled experimental evidence for the AAR's effectiveness is limited — it evolved in operational contexts, not research settings — its widespread adoption across military organizations worldwide and its subsequent adaptation by business organizations (Edmondson, 2019) suggest substantial face validity and perceived utility.

The premortem technique. Klein (2007) developed the premortem as a practical judgment-building tool. Before a project begins, the team imagines that the project has failed and then works backward to identify plausible causes of failure. This technique leverages two psychological principles: prospective hindsight (imagining that an event has already occurred makes it easier to generate explanations) and deliberate search for disconfirming evidence (which counteracts the natural tendency toward optimistic planning). The premortem has not been subjected to rigorous randomized controlled trials, but it represents a well-designed application of judgment research to practical settings.

Simulation and scenario-based training. High-fidelity simulation — in aviation, surgery, nuclear power, and military operations — provides concentrated exposure to varied, consequential situations without the risks of real-world experience. The evidence for simulation training in building technical skill (layer 2) is strong. The evidence for its effects on judgment (layer 3) is more limited but promising, particularly when simulations include unexpected complications, time pressure, and the need to make decisions with incomplete information — conditions that force the development of pattern recognition and mental simulation capacity. The key finding from simulation research is that fidelity to psychological demands matters more than fidelity to physical features. A low-tech tabletop exercise that faithfully reproduces the cognitive challenges of real decision-making may develop judgment more effectively than a high-tech simulation that looks realistic but simplifies the decision-making task (Salas, Rosen & DiazGranados, 2010).

3.4 CAN JUDGMENT BE EXPLICITLY TAUGHT?

The evidence suggests a qualified "partially." Judgment cannot be transmitted through direct instruction in the way that knowledge and procedures can. You cannot lecture someone into having good judgment. But judgment development can be accelerated through instructional designs that provide:

1. **Varied case exposure** — a rich library of situations that builds the pattern recognition base
2. **Immediate, specific feedback** — information about the accuracy and consequences of judgments that calibrates mental models
3. **Explicit articulation of reasoning** — techniques that make the judgment process visible and subject to examination (think-aloud protocols, case discussions, after-action reviews)
4. **Exposure to ambiguity and uncertainty** — problems without clear right answers that force the learner to reason through competing considerations rather than apply rules
5. **Environments with high validity** — the Kahneman-Klein condition: stable, learnable patterns that feedback can help internalize

What cannot be short-circuited is the *time* required for pattern libraries to develop. Judgment requires extensive experience, even when that experience is concentrated and structured through instructional design. The Dreyfus model's stages are not arbitrary — they reflect the genuine cognitive development that occurs as pattern recognition shifts from deliberate, analytical processing to fast, automatic processing. This development takes years in most professional domains, and there is no instructional shortcut that eliminates this requirement.

The honest assessment is that **we know more about the conditions that support judgment development than about how to directly train it**. The most effective approaches are not "judgment curricula" but institutional designs that create the conditions for judgment to develop: high-validity environments, rapid feedback loops, structured reflection, varied experience, and safety to make and examine errors.

3.5 TACIT KNOWLEDGE AND THE LIMITS OF ARTICULATION

Polanyi's (1966) concept of tacit knowledge — "we know more than we can tell" — provides the philosophical foundation for understanding why judgment resists direct instruction. Much of what experts know is tacit: embedded in their pattern recognition, their motor routines, their intuitive sense of what matters in a situation. Ask an experienced clinician how they recognized that a patient was deteriorating, and they will often say something like "I just knew something was wrong" — a response that is informative about the nature of expert knowledge but useless as an instructional prescription.

The tacit knowledge problem operates at every level of the competence stack but becomes dominant at layer 3. At layers 1–2, most knowledge and skill can be made explicit — articulated in textbooks, demonstrated in worked examples, practiced through structured exercises. The proportion of tacit knowledge increases as we move up the stack. At layer 3, a substantial portion of what the expert knows is tacit — it is encoded in pattern recognition systems that operate below the level of conscious awareness and cannot be fully articulated even by the expert who possesses them.

This has a specific and important implication: **the primary mechanism for transmitting judgment is not instruction but apprenticeship**. The apprentice does not learn judgment by being told how to judge; they learn it by working alongside someone who judges well, in the same

environment, on the same problems, over an extended period. The mechanisms through which this learning occurs — observation, imitation, osmosis, guided attention, shared practice — are not well-understood by cognitive science, partly because they resist the kind of controlled experimental study that produces clean findings. But the fact that they work is attested by thousands of years of professional education across cultures and domains.

The challenge for Applied Pedagogy is that apprenticeship is expensive, difficult to scale, and dependent on the quality of the master. Modern educational systems have sought to replace apprenticeship with scalable instruction, and they have largely succeeded at layers 1–2. At layer 3, the replacement has been much less successful. Case-based instruction, simulation, and structured experience are attempts to capture some of the judgment-developing power of apprenticeship in more scalable forms, but they inevitably lose something in the translation — particularly the embodied, situated, tacit dimensions of expert judgment.

3.6 CREW RESOURCE MANAGEMENT: A CASE STUDY IN JUDGMENT TRAINING

One of the most extensively studied judgment-training programs is Crew Resource Management (CRM) in aviation. CRM was developed in response to a series of aviation accidents in the 1970s and 1980s that were caused not by technical failures but by failures of communication, decision-making, and team coordination — failures of judgment, in competence stack terms.

CRM training focuses on situational awareness (perceiving and interpreting the current situation accurately), communication (sharing information and concerns clearly, regardless of hierarchy), decision-making (structured approaches to high-stakes decisions under uncertainty), and team management (coordinating multiple people toward a shared goal). It uses simulation extensively — cockpit simulators that present crews with complex, time-pressured scenarios requiring coordinated judgment.

The evidence for CRM's effectiveness is positive but not as strong as its widespread adoption might suggest. Salas et al. (2006) conducted a meta-analysis and found that CRM training improved attitudes and knowledge about teamwork and communication, with moderate effect sizes. Its effects on actual behavior in operational settings were harder to measure and less consistently documented. The strongest evidence comes from accident statistics: the rate of crew-caused accidents in commercial aviation has declined dramatically since CRM training became widespread, though it is difficult to disentangle the effects of CRM from other concurrent improvements in technology, regulation, and safety culture.

CRM's relevance to the competence stack lies less in its specific techniques than in its underlying model: judgment failures in complex domains are often not individual cognitive failures but failures of the social and organizational systems in which individuals operate. A pilot who makes a poor decision because they did not incorporate information available to the copilot has not failed individually — the team's communication system has failed. This connects directly to the environmental dimension: the structure of the team and its communication norms either enable or constrain the quality of individual judgment.

4

LAYER 4: METACOGNITION — KNOWING WHAT YOU KNOW AND DON'T KNOW

Metacognition — awareness of one's own cognitive processes, including the ability to monitor one's performance, recognize the boundaries of one's knowledge, and adjust strategies in response to difficulty — is the layer of the competence stack with the strongest independent research base after knowledge and skill. It is also the layer most directly relevant to the Dunning-Kruger problem: the observation that the skills needed to produce competent performance are the same skills needed to recognize competent performance, making metacognitive deficits self-concealing.

4.1 THE METACOGNITION FRAMEWORK

Flavell (1979) introduced the term "metacognition" to describe "knowledge and cognition about cognitive phenomena" — thinking about thinking. Subsequent research has distinguished two components:

Metacognitive knowledge — what people know about cognition in general and their own cognition in particular. This includes knowledge of effective learning strategies (knowing that spaced practice is more effective than cramming), knowledge of task demands (knowing that a logic problem requires a different approach than a reading comprehension task), and knowledge of one's own strengths and weaknesses (knowing that one is good at spatial reasoning but weak at verbal recall). Metacognitive knowledge can be directly taught and is the easiest component of metacognition to improve through instruction.

Metacognitive regulation — the processes of planning, monitoring, and evaluating one's own cognitive performance. Before beginning a task, the metacognitively skilled learner plans an approach, estimates how long it will take, and identifies potential difficulties. During the task, they monitor their progress, notice when comprehension breaks down, and adjust strategies accordingly. After the task, they evaluate the outcome and update their approaches for future tasks. Metacognitive regulation is harder to teach than metacognitive knowledge because it requires the learner to engage in real-time monitoring of cognitive processes that are, by their nature, partially opaque.

4.2 DESIRABLE DIFFICULTIES AND THE METACOGNITION PARADOX

Before examining the evidence for metacognitive training, it is important to understand a paradox that pervades this entire domain: the conditions that produce the best learning often feel like the worst learning, and the conditions that feel like the best learning often produce the worst.

Robert Bjork (1994) introduced the concept of "desirable difficulties" — learning conditions that make initial performance worse but improve long-term retention and transfer. Spacing practice (rather than massing it), interleaving different problem types (rather than blocking them), testing (rather than re-studying), and generating answers before being shown them (rather than simply studying them) all make learning harder in the moment and produce more errors during practice — but they produce substantially better learning outcomes over time. The key word is "desirable":

these difficulties are beneficial precisely because they force deeper cognitive processing and prevent the learner from relying on shallow, short-term cues that create an illusion of mastery.

The metacognition paradox is that learners consistently misjudge which conditions are helping them learn. Deslauriers et al. (2019), in a study of Harvard physics courses (1,241 citations), directly demonstrated this: students in active learning sections performed better on tests but reported *lower* satisfaction and *lower* perceived learning than students in passive lecture sections. The students experiencing the most effective instruction felt they were learning less because the instruction was harder. This is a metacognitive failure of the first order — learners are using "ease of processing" as a cue for learning quality, and the cue is systematically misleading.

This has profound implications for the competence stack. If metacognition at layer 4 is supposed to enable learners to monitor their own learning effectively, but learners' metacognitive cues are systematically miscalibrated — pointing them away from the most effective learning conditions and toward the least effective ones — then metacognitive training is not merely a nice addition to instruction. It is a *prerequisite* for learners to accept and engage with the kinds of instruction (productive failure, desirable difficulties, spaced practice) that produce the best outcomes at layers 1–3. Without metacognitive calibration, learners will resist the very instructional designs that would most effectively develop their competence.

The Bjork framework also introduces the distinction between **storage strength** and **retrieval strength** in memory. Storage strength reflects how deeply something has been encoded; retrieval strength reflects how easily it can be accessed at a given moment. Desirable difficulties work by temporarily reducing retrieval strength (making things harder to recall right now) while increasing storage strength (making the memory more durable and more richly connected to other knowledge). From a metacognitive perspective, learners typically monitor retrieval strength — "can I recall this easily?" — and mistake it for storage strength — "have I learned this well?" This is why students who cram before exams feel prepared (high retrieval strength) but forget everything within weeks (low storage strength). Teaching learners to distinguish these two types of memory strength — and to trust storage-building practices even when they feel difficult — is a specific, trainable metacognitive skill.

4.3 THE EVIDENCE FOR METACOGNITIVE TRAINING

The evidence that metacognitive training improves learning outcomes is substantial, though the mechanisms and effect sizes are debated.

Self-explanation training. Chi et al. (1989) demonstrated that prompting learners to explain material to themselves as they study it — rather than simply reading it — substantially improves comprehension and transfer. The self-explanation effect has been replicated extensively and works through two mechanisms: it forces the learner to identify gaps in their understanding (metacognitive monitoring) and it promotes the active integration of new information with prior knowledge (deeper processing). Bisra et al. (2018) conducted a meta-analysis of self-explanation research and found a moderate positive effect ($d \approx 0.5$) on learning outcomes, with stronger effects when self-explanation is prompted (rather than spontaneous) and when it is combined with other active learning strategies. The effect is robust across ages and domains.

Calibration training. People are generally overconfident in their knowledge — they believe they know more than they actually do. This miscalibration is not randomly distributed; it is worst among the least knowledgeable, which creates the Dunning-Kruger dynamic discussed below. Can calibration be trained? The evidence says yes, with important caveats. Carpenter et al. (2018), in a study with 168 citations, demonstrated that domain-general metacognitive ability — specifically,

the capacity to distinguish between correct and incorrect judgments — can be enhanced through adaptive training. Participants who completed a metacognitive training program showed improved discrimination between confident-and-correct versus confident-and-incorrect judgments, and this improvement transferred to untrained domains.

The calibration training literature more broadly suggests that two approaches are effective. **Outcome feedback** — telling people after each judgment whether they were right or wrong — improves calibration slowly over many trials. **Process feedback** — helping people understand *why* they were overconfident, what cues they relied on, and what cues they should have attended to — improves calibration more rapidly and durably. Dunlosky and Rawson (2012) found that overconfidence is a specific predictor of underachievement: students who believe they have mastered material when they have not are less likely to engage in additional study, producing a self-reinforcing cycle of miscalibration and poor performance. This has direct implications for the competence stack: miscalibrated confidence at layer 4 actively undermines competence at all other layers, because the learner does not know what they need to improve and therefore does not improve it.

Prediction-first pedagogy. A particularly promising approach to developing metacognition is what might be called prediction-first pedagogy — instructional designs that require learners to make predictions before receiving information. When a learner predicts what will happen in an experiment, or estimates the answer to a problem, or guesses what a new concept means before it is explained, they are engaging in metacognitive activity: they must assess what they already know, generate hypotheses, and commit to specific expectations. When the prediction is then compared to the actual outcome, the discrepancy between prediction and reality provides a powerful metacognitive signal — it makes knowledge gaps visible and concrete.

Kapur's productive failure paradigm (Kapur, 2008; Kapur, 2024) is the most rigorously studied version of this approach. Productive failure deliberately places learners in situations where they must attempt to solve problems before receiving instruction. The initial attempts typically fail — students generate multiple solution approaches, most of which are incorrect or incomplete. But this failure serves four functions that Kapur describes as the 4A model: **Activation** of prior knowledge (struggle forces learners to mobilize everything they already know), **Awareness** of gaps (the failure makes knowledge deficits concrete and specific), **Affect** (the emotional engagement of struggling with a genuine problem), and **Assembly** (the subsequent instruction has richer material to organize, because learners now have activated prior knowledge and identified specific gaps that instruction can address).

The evidence for productive failure is strong and growing. A meta-analysis of over 50 studies (160+ comparisons), as reported in Kapur's 2024 book, found that productive failure produces equivalent or slightly superior procedural skill compared to direct instruction, and substantially superior conceptual understanding and transfer — with effect sizes as large as two academic years of learning for transfer outcomes. These findings have been replicated by independent researchers across mathematics, science, and other domains, and a 2023 review by leading scholars concluded that deep learning comes from combining inquiry with direct instruction, which is precisely what productive failure does.

The metacognitive significance of productive failure is underappreciated. By requiring learners to attempt problems before instruction, it forces them to confront the limits of their current understanding — to experience, concretely and specifically, what they do not know. This is metacognitive training embedded in content instruction, rather than bolted on as a separate activity.

Reflection and self-assessment. The broader literature on reflection as a metacognitive tool is more mixed. Structured reflection — guided by specific prompts, focused on concrete experiences, and connected to future action — can improve metacognitive monitoring and learning strategy use. Unstructured reflection — "journal about what you learned today" — often becomes performative,

producing text that satisfies the assignment requirement without engaging genuine metacognitive processes. Andrade (2019), in a critical review of self-assessment research (492 citations), found that self-assessment improves learning outcomes when students are given clear criteria against which to evaluate their work, but not when self-assessment is vague or unfocused. This aligns with the broader finding in the assessment literature (L1-003) that the specificity and actionability of feedback — including self-generated feedback — is what determines its effectiveness.

4.4 THE DUNNING-KRUGER EFFECT AND ITS IMPLICATIONS

The Dunning-Kruger effect (Kruger & Dunning, 1999; 6,680 citations, FWCI 51.86) is perhaps the most important finding in all of metacognition research for the competence stack. Kruger and Dunning demonstrated that people who perform poorly on tests of logical reasoning, grammar, and humor are disproportionately likely to overestimate their own performance — not just slightly, but dramatically. Participants in the bottom quartile of performance estimated themselves to be above average. Moreover, this overconfidence was specifically linked to metacognitive deficits: the same skills required to produce correct answers (logical reasoning, grammatical knowledge) were the same skills required to recognize that one's answers were wrong.

This creates a vicious cycle. The person who lacks competence also lacks the metacognitive capacity to recognize their incompetence, which means they have no motivation to improve, which means their incompetence persists, which means their metacognitive deficit persists. Dunning (2011) described this as "the anosognosia of everyday life" — a parallel to the neurological condition in which brain-damaged patients are unaware of their deficits.

The finding has been extensively replicated but also extensively debated. A significant criticism is that part of the observed effect is statistical artifact — regression to the mean combined with floor and ceiling effects on performance scales (Krueger & Mueller, 2002). When performance is poor, there is more room for overestimation than underestimation, producing asymmetric errors that can look like a systematic metacognitive deficit even in the absence of one. Dunning and colleagues have responded by showing that the effect persists even after controlling for statistical artifacts and that it has specific mechanistic explanations rooted in metacognitive monitoring failures (Dunning, Johnson, Ehrlinger & Kruger, 2003).

For the competence stack, the important question is not whether the effect is real — it is — but whether it can be remediated. Can metacognitive training help those who most need it? The evidence here is cautiously optimistic but limited.

Kruger and Dunning (1999) demonstrated that training in logical reasoning improved both performance *and* the accuracy of self-assessment — participants who received training became better at the task and also more accurate in estimating their own ability. This suggests that the metacognitive deficit is not fixed but is responsive to skill development. As competence increases, the metacognitive capacity to recognize competence increases as well. The implication is that metacognitive training may not need to be separate from content training — building competence at layers 1–2 may naturally improve metacognitive accuracy at layer 4.

However, there is a bootstrap problem. If the people who most need metacognitive training are precisely those who are least likely to recognize their need for it, how do you get them to engage with the training in the first place? Productive failure may offer one answer: by placing learners in situations where their incompetence becomes concretely, unavoidably visible — not through an authority telling them they are wrong, but through their own experience of being unable to solve a problem — the metacognitive gap becomes salient in a way that cannot easily be ignored or rationalized away.

4.5 WHEN REFLECTION HELPS VS. HURTS

Not all metacognitive activity is beneficial. Self-reflection can become **ruminative** — repetitive, self-focused thinking that does not lead to new insights or adaptive action — particularly for individuals with anxiety, depression, or low self-esteem. Nolen-Hoeksema (2000) demonstrated that rumination exacerbates negative mood and impairs problem-solving, which means that reflexive prompts to "reflect on your learning" can be actively harmful for some learners.

Self-reflection can also become **performative** — producing text or behaviors that mimic genuine metacognitive activity without actually engaging it. When reflection is required for a grade or a course requirement, students learn to produce reflection-shaped artifacts that satisfy institutional expectations without involving genuine self-examination. This is a specific instance of Goodhart's Law: when a measure becomes a target, it ceases to be a good measure.

The conditions under which reflection genuinely improves learning are fairly well-specified:

1. **Specificity** — reflection focused on concrete, recent experiences rather than abstract generalities
2. **Action-orientation** — reflection connected to specific future actions ("next time I will…") rather than purely retrospective
3. **Safety** — an environment where honest self-assessment is not punished (connecting directly to layer 5 and the environmental dimension)
4. **Structure** — guided by prompts, criteria, or frameworks rather than left entirely open
5. **Accuracy-focused** — emphasis on calibrating self-assessment accuracy rather than on feeling good about one's performance

The research suggests that reflection is a tool, not a virtue — valuable when well-designed and potentially harmful when poorly designed or mandated without regard for individual differences.

4.6 SELF-REGULATED LEARNING: THE META-SKILL THAT CONNECTS THE LAYERS

Self-regulated learning (SRL) — the capacity to plan, monitor, and evaluate one's own learning processes — sits at the intersection of metacognition (layer 4) and skill development (layer 2). The L1-002 investigation identified self-regulation as one of the strongest predictors of long-term educational outcomes, citing Moffitt et al.'s (2011) Dunedin study, which found that childhood self-control predicted adult outcomes across physical health, financial stability, and social functioning in a continuous gradient even after controlling for IQ and social class.

Panadero (2017), in a comprehensive review of six major SRL models (2,305 citations), found convergence on three core phases — forethought (planning, goal-setting), performance (monitoring, strategy deployment), and self-reflection (evaluation, attribution) — but divergence on the role of motivation and emotion. For the competence stack, the important finding is that SRL is not a single capacity but a coordinated deployment of multiple capacities: metacognitive monitoring (layer 4), strategic knowledge (layer 1), procedural skill (layer 2), and motivational regulation (connected to layer 5 through persistence and tolerance for difficulty).

The L1-002 investigation established that self-regulation can be taught through direct instruction — and that direct instruction in self-regulation strategies is more effective than indirect approaches that attempt to activate self-regulation through environmental cues alone. This finding has important implications for the competence stack. It suggests that at least some aspects of the upper layers *can* be directly taught, not just environmentally facilitated. The key distinction may be

between the *skills* of self-regulation (which are teachable through instruction) and the *disposition* to deploy those skills consistently (which may depend more on environment and character). A student can be taught the strategy of monitoring their own comprehension while reading — but whether they consistently do so, especially when it is difficult or when no one is watching, may depend on the motivational and character dimensions that direct instruction alone cannot guarantee.

The connection between SRL and the Dunning-Kruger problem is particularly important. Effective self-regulation requires accurate metacognitive monitoring — you cannot regulate your learning effectively if you cannot tell whether you are learning effectively. This means that the Dunning-Kruger deficit is not just a metacognitive problem; it is a self-regulation problem. Learners who overestimate their competence do not merely feel good about bad performance — they fail to deploy the regulatory strategies (additional study, strategy change, help-seeking) that would correct their deficits. The vicious cycle is metacognitive-regulatory, not merely metacognitive: poor monitoring leads to poor regulation, which leads to poor learning, which perpetuates poor monitoring.

Breaking this cycle requires interventions that address both components: metacognitive monitoring (helping learners see their actual performance accurately) and regulatory strategy (giving them effective tools to use once they recognize the gap). Productive failure addresses the monitoring side by making gaps experientially visible. Strategy instruction addresses the regulatory side by providing concrete tools. The combination — experiencing your gaps, then learning what to do about them — is more powerful than either component alone.

5 LAYER 5: CHARACTER, DISPOSITION, AND EPISTEMIC VIRTUE

5.1 THE THINNEST EVIDENCE, THE HIGHEST STAKES

Layer 5 of the competence stack — character and disposition — is where the evidence base is thinnest and the claims are most ambitious. COMPETENCE-TARGET.md identifies this layer as "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." It further claims that these are "not fixed personality traits but epistemic and moral dispositions that can be cultivated or degraded by environment." This section examines what the evidence supports.

The relevant literatures include intellectual humility (a construct that has received significant empirical attention in the last decade), psychological safety (Edmondson's work on team learning), virtue epistemology (the philosophical tradition concerned with intellectual character), and the small but important literature on character education and moral development.

5.2 INTELLECTUAL HUMILITY: WHAT THE EVIDENCE SHOWS

Intellectual humility — roughly, the recognition of one's own cognitive fallibility and the disposition to own one's intellectual limitations — has gone from a primarily philosophical construct to an active empirical research area in less than a decade. Porter et al. (2022), in a comprehensive review published in *Nature Reviews Psychology* (FWCI 40.81), synthesized the rapidly growing literature and reached several important conclusions.

Intellectual humility can be measured with acceptable reliability. Multiple validated scales exist, including those developed by Leary et al. (2017) and Alfano et al. (2017). Factor analyses consistently identify intellectual humility as distinct from general humility, agreeableness, openness to experience, and other related constructs. It is not simply "nice personality" by another name. The construct has both intrapersonal dimensions (accurate self-assessment, recognition of fallibility) and interpersonal dimensions (respectful engagement with opposing views, willingness to revise beliefs in response to evidence).

Intellectual humility is associated with better epistemic outcomes. People who score higher on intellectual humility measures show greater openness to new information, more willingness to revise beliefs in response to evidence, better discrimination between strong and weak arguments regardless of whether they align with the person's prior beliefs, reduced susceptibility to belief-perseverance biases, and more accurate self-assessment of knowledge (Porter et al., 2022). These associations are moderate in size (correlations typically in the $r = 0.2–0.4$ range) and have been replicated across multiple studies and measurement instruments.

Intellectual humility is somewhat stable but also responsive to context. Trait intellectual humility shows moderate test-retest reliability — it is more stable than a mood but less stable than a personality trait. Critically, context matters: people display more intellectual humility in domains where they feel secure and less in domains where their identity is threatened. This suggests that intellectual humility is not purely a trait to be cultivated but partly a state that is facilitated or inhibited by environmental conditions.

The evidence that intellectual humility can be directly trained is thin. This is the critical finding for Applied Pedagogy. Despite the growing empirical literature on measurement and correlates, there are almost no rigorous intervention studies demonstrating that intellectual humility can be increased through training. Porter et al. (2022) note this gap explicitly: "a critical next step for research is to develop and test IH interventions." Some preliminary evidence suggests that perspective-taking exercises, exposure to disagreement from trusted sources, and environments that model intellectual humility can increase state-level intellectual humility, but the durability and transferability of these effects are unknown.

Porter and Schumann (2018) found that brief interventions — such as reading about the incremental (vs. fixed) nature of intelligence — can increase self-reported intellectual humility and willingness to engage with opposing views. But the long-term effects of such interventions are unknown, and self-reported intellectual humility may not track behavioral intellectual humility accurately (the same metacognitive problems that affect other forms of self-assessment likely apply here).

5.3 THE PHILOSOPHICAL FOUNDATIONS: VIRTUE EPISTEMOLOGY

The philosophical literature on intellectual virtue provides the conceptual framework within which the empirical work operates, even when the empirical researchers do not explicitly acknowledge it.

Whitcomb, Battaly, Baehr, and Howard-Snyder (2017) defined intellectual humility as "a disposition to own one's cognitive limitations" — not low self-esteem or false modesty, but accurate recognition of what one does and does not know. Church and Samuelson (2017) provided a broader introduction to the philosophy and science of intellectual humility, identifying it as a "master virtue" that enables other intellectual virtues: you cannot be open-minded, intellectually courageous, or appropriately cautious without first being honest about the limits of your own knowledge.

The virtue epistemology tradition (Baehr, 2011; Zagzebski, 1996) argues that intellectual character — the cluster of dispositions that includes intellectual honesty, intellectual courage, intellectual humility, intellectual perseverance, and intellectual autonomy — is cultivated through practice, much like moral character in the Aristotelian tradition. On this view, one becomes intellectually honest not by being told to be honest but by repeatedly engaging in honest inquiry in contexts that reward honesty. This "practice makes character" model has deep philosophical roots but limited empirical testing.

The gap between the philosophical literature and the empirical literature is significant. Philosophers have provided rich, detailed accounts of what intellectual virtues are and why they matter. Psychologists have developed measures and identified correlates. But almost no one has rigorously tested whether and how intellectual virtues can be cultivated through educational interventions. This is the most important gap in the competence formation literature.

5.4 PSYCHOLOGICAL SAFETY: THE ENVIRONMENTAL CONDITION

Amy Edmondson's work on psychological safety is the most extensively researched aspect of the environmental dimension of the competence stack. Her 1999 paper (10,040 citations, FWCI 42.93) demonstrated that teams with higher psychological safety — defined as "a shared belief held by members of a team that the team is safe for interpersonal risk taking" — showed more learning behavior: asking questions, reporting errors, experimenting with new approaches, and requesting feedback. The effect was not merely that psychologically safe teams felt better. They *learned* more

effectively because they engaged in behaviors essential to learning that psychologically unsafe teams suppressed.

Edmondson's subsequent work, particularly *The Fearless Organization* (2019), extended these findings across industries and identified specific leadership behaviors that create or destroy psychological safety: framing work as a learning problem rather than an execution problem; acknowledging one's own fallibility; modeling curiosity; responding productively to bad news, questions, and mistakes; and explicitly sanctioning interpersonal risk-taking. She documented how the absence of psychological safety — particularly in hierarchical organizations — leads to specific, predictable pathologies: avoidable errors are not reported and therefore not corrected; problems are concealed until they become crises; employees learn to suppress honest assessment and produce the appearance of confidence; and organizational learning ceases even as individual activity continues.

The relevance to the competence stack is direct and profound. COMPETENCE-TARGET.md's diagnostic question 5 — "Are they allowed to tell the truth in that system?" — is essentially a restatement of Edmondson's construct. And the document's claim that this dimension is multiplicative, not additive, finds support in Edmondson's research: psychological safety does not just add to learning; in its absence, the other conditions for learning (feedback, error correction, honest self-assessment) cease to function. An environment without psychological safety does not merely slow competence development. Over time, it actively degrades the upper layers of the competence stack by severing the feedback loops through which judgment, metacognition, and character are maintained.

5.5 ERROR MANAGEMENT CULTURE: THE ORGANIZATIONAL COMPLEMENT

van Dyck, Frese, Baer, and Sonnentag (2005) introduced the concept of "error management culture" — an organizational culture that distinguishes between error *prevention* (trying to minimize errors, which is necessary but insufficient) and error *management* (having processes for detecting, analyzing, and learning from errors when they inevitably occur). Their two-study replication (795 citations) demonstrated that organizations with stronger error management cultures had better financial performance, controlling for industry, size, and age. The mechanism they identified was exactly what the competence stack predicts: in error management cultures, errors are detected sooner, analyzed more thoroughly, and used as information for systemic improvement rather than as grounds for blame.

The error management culture concept connects directly to Argyris's (1977) distinction between single-loop and double-loop learning. **Single-loop learning** corrects errors within existing frameworks — doing the same thing more accurately. **Double-loop learning** examines and revises the frameworks themselves — questioning the assumptions that produced the error. Argyris argued that most organizations are systematically biased toward single-loop learning because double-loop learning requires people to question their own mental models, expose their reasoning to scrutiny, and admit that their fundamental assumptions may be wrong. This is psychologically threatening, particularly in environments that penalize error, which means that the very organizations that most need double-loop learning are those whose cultures make it most difficult.

Senge (1990), in *The Fifth Discipline*, extended this analysis to argue that organizational learning requires five disciplines: personal mastery, mental models, shared vision, team learning, and systems thinking. The relevant discipline for the competence stack is "mental models" — the deeply ingrained assumptions, generalizations, and images that influence how people understand the world and take action. Senge argued that effective organizational learning requires people to

surface and test their mental models — to make their thinking visible and subject it to collective scrutiny. This connects directly to metacognition (layer 4) and intellectual humility (layer 5): the organizational disciplines Senge describes are essentially institutional practices for developing and maintaining the upper layers of the competence stack at the organizational level.

5.6 CAN EPISTEMIC CHARACTER BE CULTIVATED THROUGH EDUCATIONAL DESIGN?

The honest answer is: **we don't know, but the evidence points toward environmental design rather than direct training**.

The intellectual humility literature suggests that the construct is real, measurable, and associated with better epistemic outcomes — but there are almost no rigorous intervention studies. The psychological safety literature demonstrates that environments can either support or suppress the behaviors associated with epistemic character — but it does not directly demonstrate that environmentally supported behavior translates into durable character traits. The virtue epistemology tradition provides compelling philosophical arguments for why character should be cultivatable through practice — but these arguments remain largely untested empirically.

What the evidence does support is a two-part model:

Part 1: Environmental design is the primary intervention. Creating environments that reward intellectual honesty, treat error as information, tolerate uncertainty, model intellectual humility from positions of authority, and protect people from social punishment for admitting ignorance — this is the most evidence-supported approach to developing epistemic character. The evidence from Edmondson, van Dyck, and Argyris all points in the same direction: the institutional environment is not merely a context for individual development but a first-order determinant of whether the upper layers of competence develop or degrade.

Part 2: Explicit modeling and practice may help, but the evidence is preliminary. Exposure to intellectual humility from respected authorities (teachers, mentors, leaders), structured practice in perspective-taking and evidence evaluation, and environments where intellectual character is visibly valued and rewarded — these are plausible candidates for character-building interventions. But the empirical evidence for their effectiveness is thin, consisting primarily of correlational studies and short-term intervention studies with unknown long-term effects.

5.7 CULTURE AND CHARACTER FORMATION

The cross-cultural dimension of character formation is important but underresearched. Different cultures have different norms around admitting error, expressing uncertainty, and challenging authority. In high power-distance cultures (Hofstede, 2001), challenging a superior's judgment is socially costly regardless of its epistemic merit. In shame cultures, admitting error may carry reputational costs that far exceed the informational value of the admission. These cultural factors interact with the competence stack in complex ways that the predominantly Western research literature has barely begun to address.

What little cross-cultural evidence exists suggests that the *psychological mechanisms* underlying competence formation are broadly similar across cultures — people everywhere benefit from feedback, learning environments, and accurate self-assessment — but the *social contexts* in which these mechanisms operate vary enormously. Creating psychological safety in a hierarchical Confucian educational context requires different strategies than creating it in an egalitarian Nordic one, even

though the underlying goal — enabling people to admit error and pursue truth without social punishment — is the same.

Honor cultures, in which reputation is paramount and any sign of weakness is exploited, present particular challenges for the competence stack. In honor cultures, the behaviors that the competence stack requires at layers 4 and 5 — admitting ignorance, expressing uncertainty, changing one's mind in response to evidence — are precisely the behaviors that carry the highest social costs. Educational institutions embedded in honor cultures face a structural tension between the epistemic requirements of competence and the social requirements of reputation maintenance. This tension is underresearched and may represent one of the most important L2 investigations for Applied Pedagogy.

Part III

ENVIRONMENT AND TIME

6

THE ENVIRONMENTAL DIMENSION: HOW INSTITUTIONS PROMOTE OR DEGRADE COMPETENCE

6.1 THE MOTIVATION CONNECTION: SDT AND THE UPPER LAYERS

Before examining the environmental evidence directly, it is worth connecting to the L1-002 investigation's findings on self-determination theory. SDT identified three basic psychological needs — autonomy, competence, and relatedness — whose satisfaction is essential for high-quality motivation and learning. The L1-002 investigation documented robust evidence that controlling environments (those that use rewards, punishments, surveillance, and evaluative pressure to direct behavior) undermine intrinsic motivation and produce shallower engagement.

The connection to the competence stack is deeper than it may initially appear. SDT's "controlling vs. autonomy-supportive" distinction maps directly onto the environmental conditions that promote or suppress upper-layer competence development. A controlling environment — one that rewards correct performance and punishes error — is precisely the environment in which metacognitive honesty (admitting what you don't know) and intellectual humility (revising your beliefs) carry the highest costs. An autonomy-supportive environment — one that provides rationale, acknowledges the learner's perspective, and minimizes evaluative pressure — is precisely the environment in which the upper layers can develop, because the learner is free to acknowledge gaps, make mistakes, and update their understanding without social penalty.

The L1-002 finding that intrinsic motivation declines systematically from elementary through secondary school — and that this decline correlates with increasing institutional control — takes on additional significance in the competence stack framework. The same institutional features that suppress motivation may also suppress the development of judgment, metacognition, and character. The motivational decline documented by L1-002 may be only the most visible symptom of a broader competence-formation failure: schools that become more controlling as students advance may be systematically preventing the development of the very capacities that education is supposed to produce at its highest aspirations.

This convergence between SDT (L1-002), the assessment paradox (L1-003), and the environmental dimension of the competence stack (this investigation) constitutes one of the strongest cross-cutting findings in the lab's work to date. Controlling, high-stakes, evaluatively pressured educational environments do not merely reduce motivation and distort assessment — they structurally prevent the development of full-stack competence. The three problems are not separate; they are the same problem viewed from three different angles.

6.2 THE MULTIPLICATIVE CLAIM

COMPETENCE-TARGET.md makes a strong claim: the environmental dimension — whether people are allowed to tell the truth in a system — is not merely additive but **multiplicative**. An environment that penalizes honesty does not just suppress truth-telling; over time it degrades the capacity to *perceive* truth. What does the evidence support?

6.3 HOW TOXIC ENVIRONMENTS MANUFACTURE INCOMPETENCE

The strongest evidence for the "manufacture of incompetence" comes from healthcare and aviation safety research, where the consequences of suppressed error reporting are measurable and severe.

Healthcare. Edmondson (1996, 2004) documented how hospitals with punitive error-reporting cultures experienced not fewer reported errors but fewer *detected* errors — and more patient harm. When nurses and residents feared punishment for reporting medication errors, they stopped reporting. When they stopped reporting, the errors continued but were no longer visible to the system. When the errors were no longer visible, systemic improvements could not be made. The result was not just a culture of silence but a systematic degradation of the organization's ability to perceive and respond to problems — a degradation of collective competence at layers 3–5.

Vaughan (1996), in her analysis of the Challenger space shuttle disaster, introduced the concept of "normalization of deviance" — the process by which initially concerning behaviors become accepted as normal through repeated exposure without consequence. O-ring erosion that should have been a red flag became an "acceptable risk" because previous launches with similar erosion had not resulted in catastrophe. The organizational culture at NASA had gradually degraded the capacity of individuals within it to perceive risk accurately — their layer 3 (judgment) and layer 4 (metacognition) had been systematically compromised by an institutional environment that normalized the very signals they should have been attending to.

The feedback loop model. Drawing on these and similar cases, the evidence supports the following model of how toxic environments degrade competence:

1. **Suppression of honest reporting** — error signals are not transmitted because the messenger faces punishment
2. **Loss of system visibility** — without error signals, the organization (and individuals within it) lose awareness of problems
3. **Mental model degradation** — without disconfirming evidence, individuals' mental models drift from reality and are not corrected
4. **Normalization of deviance** — behaviors and standards that would previously have been flagged as problematic become the new baseline
5. **Metacognitive erosion** — individuals lose the capacity to distinguish between "things are fine" and "we have stopped seeing problems"
6. **Character adaptation** — people learn that honesty is punished and performance of confidence is rewarded, and they adapt accordingly

This model explains how organizations can contain individually competent people and yet produce collectively incompetent outcomes. The system destroys competence not by employing incompetent individuals but by creating conditions that prevent competent individuals from exercising — and eventually from developing — their upper-layer capacities.

6.4 HOW CONSTRUCTIVE ENVIRONMENTS BUILD COMPETENCE

The inverse is also supported: environments that create the conditions for truth-telling, error-learning, and honest self-assessment can actively promote the development of upper-layer competence.

Edmondson's research on high-performing surgical teams (2003) identified the specific leadership behaviors that enabled rapid learning during the adoption of minimally invasive cardiac surgery — a technically challenging innovation that required established teams to fundamentally

change their established practices. Teams that learned fastest had leaders who framed the innovation as a learning challenge (not an execution challenge), invited input from all team members (not just senior members), responded to problems with curiosity (not blame), and explicitly acknowledged their own fallibility ("I don't know how this will go, and I need everyone's eyes and ears"). These behaviors created the psychological safety that enabled the feedback loops necessary for rapid skill acquisition and judgment development.

The implications for education are clear but challenging to implement. If competence at layers 3–5 depends on environmental conditions — and the evidence strongly suggests it does — then educational design must attend to the learning environment as a first-order variable, not merely the content and methods of instruction. A brilliantly designed curriculum delivered in a punitive, anxiety-inducing, or status-competitive environment will fail to develop judgment, metacognition, or character, no matter how well it builds knowledge and skill.

6.5 THE STRUCTURE OF FEEDBACK LOOPS

The quality of feedback loops in an institution affects individual competence development over time through several mechanisms.

Feedback latency. The time between action and feedback profoundly affects learning. In high-feedback environments (emergency medicine, air traffic control, competitive chess), the consequences of decisions are apparent within minutes or hours, allowing rapid calibration of mental models. In low-feedback environments (policy-making, long-term investing, most educational administration), the consequences may not be apparent for months or years, making mental model calibration extremely difficult. Einhorn and Hogarth (1978) demonstrated that delay between action and outcome systematically degrades the quality of learning from experience, even when the outcome is eventually observed.

Feedback specificity. Vague feedback ("good job" or "needs improvement") provides little information for improving performance, as the L1-003 investigation documented extensively. Specific, task-focused feedback that identifies precisely what was done well or poorly and why provides the information needed for genuine improvement. In the context of the competence stack, feedback specificity matters at every layer: specific feedback on knowledge gaps (layer 1), specific feedback on skill execution (layer 2), specific feedback on judgment calls and their outcomes (layer 3), specific feedback on metacognitive accuracy ("you were confident about X but wrong; you were uncertain about Y but right") (layer 4), and visible modeling of how authority figures handle being wrong (layer 5).

Feedback attribution. How outcomes are attributed — to skill, luck, effort, system factors, or personal inadequacy — affects what is learned from them. In blame cultures, negative outcomes are attributed to individual failure, which discourages risk-taking and error reporting. In learning cultures, negative outcomes are attributed to systemic factors and information gaps, which encourages analysis and improvement. Edmondson (2019) noted that the language of attribution is a specific, observable behavior that leaders can consciously adopt: saying "what can we learn from this?" rather than "who is responsible for this?" is a small linguistic change that signals a fundamentally different organizational relationship to error.

7

COMPETENCE OVER TIME: MAINTENANCE, DECAY, AND THE PROBLEM OF STAGNATION

The competence stack is not a structure that, once built, remains stable. Competence at every layer requires active maintenance, and without it, degrades over time. The dynamics of this maintenance and degradation differ across layers in ways that matter for curriculum design.

7.1 KNOWLEDGE AND SKILL DECAY

At layers 1–2, the dynamics of decay are relatively well understood. Knowledge that is not retrieved decays according to well-characterized forgetting curves (Ebbinghaus, 1885). Skills that are not practiced atrophy. Medical knowledge degrades measurably after residency (Choudhry, Fletcher & Soumerai, 2005), surgical skills deteriorate during periods of inactivity, and language proficiency declines without use. The remedies at these layers are correspondingly straightforward: retrieval practice, spaced review, and continued use maintain competence.

7.2 JUDGMENT DRIFT

At layer 3, the dynamics are more subtle and more dangerous. Expert judgment does not simply decay through disuse — it can actively drift. As the environment changes, the patterns that experts have internalized become increasingly mismatched to current reality. An experienced physician whose training predates a major revision in diagnostic criteria may continue to apply outdated pattern libraries with the same confidence they applied current ones. An investor whose intuitions were forged in a bull market may apply those intuitions in a fundamentally different market environment. The expert does not perceive this drift because the same automaticity that makes expert judgment fast and efficient also makes it resistant to updating.

This phenomenon — expert judgment becoming *less* accurate over time as the environment changes while the expert's pattern library does not — has received less attention than it deserves. It represents a specific failure mode of the competence stack: the expert has high confidence (because the pattern recognition feels the same) but declining accuracy (because the patterns have become mismatched). This is worse than novice ignorance because the expert does not know they are wrong and has no motivation to update.

The remedy for judgment drift is systematic exposure to feedback that reveals mismatches between the expert's mental models and current reality. This requires institutional structures — ongoing case review, peer feedback, exposure to disconfirming evidence — that many professional environments fail to provide once initial training is complete. In most professions, continuing education focuses on knowledge updates (layer 1) rather than judgment recalibration (layer 3), which means that the layer most vulnerable to drift receives the least maintenance.

7.3 METACOGNITIVE COMPLACENCY

At layer 4, the risk is what might be called metacognitive complacency — the gradual decline in self-monitoring that accompanies increasing expertise and familiarity. As competence increases

and tasks become more routine, the metacognitive monitoring that was active during learning becomes less engaged. The expert who carefully monitored their own reasoning as a trainee may, years later, execute the same reasoning automatically without monitoring it at all. This is efficient when the routine reasoning is correct, but it means that errors in the routine — whether from judgment drift, changed circumstances, or simple mistakes — are less likely to be caught.

The research on automation in aviation provides a vivid parallel. As cockpit automation increased, pilots' monitoring of automated systems decreased, leading to "automation complacency" — a documented failure mode in which pilots fail to notice when automated systems malfunction because they have learned to trust them (Parasuraman & Manzey, 2010). The analogy to cognitive expertise is direct: as expert judgment becomes automated (in the cognitive sense), the metacognitive monitoring that would catch errors in that judgment declines. The result is that experts may be simultaneously more accurate *on average* and more vulnerable to *catastrophic errors* than less experienced practitioners — because when the expert's automatic processes fail, no metacognitive safety net catches the failure.

7.4 CHARACTER UNDER PRESSURE

At layer 5, the maintenance challenge is environmental. Epistemic character — intellectual honesty, courage, humility — is not a stable trait but a disposition that is continually reinforced or eroded by the environment. A person who joins an organization with strong epistemic norms will find it relatively easy to maintain and develop their epistemic character. The same person, after years in an organization that rewards confidence performance and punishes honest uncertainty, may find their epistemic character substantially degraded — not through any conscious choice but through gradual, incremental adaptation to the incentive structure.

This environmental erosion of character is the most insidious form of competence degradation because it is slow, invisible to the person experiencing it, and self-reinforcing. The person who has adapted to a toxic environment by suppressing their honest assessments does not experience themselves as having lost something — they experience themselves as having become more "realistic" or "pragmatic." The epistemic character has not been destroyed; it has been buried under layers of adaptive behavior, and whether it can be recovered — and on what timescale — is an open empirical question.

The implication for curriculum design is that building competence at layer 5 is not a one-time educational achievement but an ongoing environmental requirement. Applied Pedagogy cannot simply graduate people with epistemic character and send them into the world; it must prepare them to recognize and resist the environmental forces that will erode that character, and it must advocate for the institutional conditions that sustain it.

Part IV

ASSESSMENT AND SYNTHESIS

8 ASSESSMENT OF LAYERS 3–5: CAN WE MEASURE WHAT MATTERS?

The practical question of whether judgment, metacognition, and character can be assessed with acceptable validity and reliability is closely related to the L1-003 investigation, but raises distinct challenges that the assessment literature has only partially addressed.

8.1 ASSESSING JUDGMENT (LAYER 3)

The most established approaches to assessing judgment come from professional certification:

Standardized patient encounters in medical education present trainees with actors portraying patients and evaluate their diagnostic reasoning, clinical decision-making, and communication. These assessments have moderate reliability (typically r = 0.5–0.7 across raters) and reasonable validity for predicting subsequent clinical performance, though they are expensive and time-consuming to administer. Their key strength is that they assess judgment in context — the trainee must integrate knowledge, skills, and judgment simultaneously, as they would in practice.

Situational judgment tests (SJTs) present written or video scenarios describing complex situations and ask respondents to rate or rank possible responses. SJTs have been shown to have incremental validity over cognitive ability tests for predicting job performance, particularly in domains that require judgment (Lievens, Buyse & Sackett, 2005). However, they are subject to coaching effects and may measure knowledge of what good judgment looks like rather than actual judgment capacity — a distinction that matters enormously for the competence stack.

The fundamental challenge of assessing judgment is that judgment, by definition, is required in situations where the "right answer" is uncertain, context-dependent, or unknown at the time the judgment is made. This means that judgment assessment cannot simply check answers against a key. It must evaluate the quality of reasoning, the appropriateness of the response to the situation's specific features, and the calibration of the decision-maker's confidence — all of which require expert evaluators, which limits scalability.

8.2 ASSESSING METACOGNITION (LAYER 4)

Metacognitive assessment is somewhat more tractable because the key variable — calibration accuracy — can be measured quantitatively.

Confidence judgments — asking learners to rate their confidence in each answer on a test — can be compared to actual accuracy to produce a calibration index. Well-calibrated learners are confident when right and uncertain when wrong; poorly calibrated learners show high confidence regardless of accuracy. This approach has been used extensively in research and has reasonable psychometric properties, though it adds time to assessments and may itself change metacognitive behavior (a form of reactivity).

Judgment of learning (JOL) accuracy — asking learners to predict their performance on an upcoming test — provides a different window into metacognitive monitoring. Comparing predicted to actual performance reveals systematic biases (overconfidence, underconfidence) and their domain specificity. Nelson and Narens (1990) established the framework for studying JOL accuracy that has been used in hundreds of subsequent studies.

Think-aloud protocols — asking learners to verbalize their thinking as they work through problems — can reveal metacognitive monitoring processes in real time. These are rich but labor-intensive to analyze and are subject to reactivity effects (the act of verbalizing may change the cognitive process being observed).

8.3 ASSESSING CHARACTER (LAYER 5)

This is where assessment becomes most challenging and most controversial.

Self-report measures of intellectual humility (Leary et al., 2017; Alfano et al., 2017) have adequate psychometric properties but face the fundamental limitation that self-report measures of self-knowledge may not accurately capture the construct they are measuring — particularly for the very individuals whose intellectual humility is lowest. This is the Dunning-Kruger problem applied to character assessment: the people who most lack intellectual humility may be least able to accurately report their level of it.

Behavioral measures — observing how people respond to disagreement, correction, and evidence that contradicts their beliefs — are more valid in principle but much harder to standardize and scale. Some researchers have developed experimental paradigms that expose participants to belief-threatening evidence and measure their updating behavior, but these are primarily research tools, not practical assessment instruments.

The assessment dilemma at layer 5 is that the most important aspects of epistemic character — honesty in the face of social pressure, courage to dissent, willingness to admit error — are precisely the aspects that are most distorted by assessment contexts. When people know they are being evaluated, they are more likely to perform intellectual humility than to actually exercise it. This is a specific instance of the broader problem that the L1-003 investigation identified: assessment tends to measure what people can perform, not what they have internalized.

The most honest assessment of where we stand on measuring layers 3–5 is this: we have promising approaches for judgment and metacognition assessment, though they are less reliable and scalable than knowledge and skill assessments; and we have barely begun to develop valid measures of epistemic character that work at scale. The gap between what we can measure at layers 1–2 and what we can measure at layers 3–5 is enormous, and this gap itself constrains educational practice. Institutions tend to optimize for what they can measure, which means they tend to optimize for layers 1–2 and neglect the upper layers — not because they don't value them but because they cannot assess them with the precision and efficiency that institutional accountability demands.

9 SYNTHESIS: WHAT A CURRICULUM DESIGNER NEEDS TO KNOW

9.1 THE EVIDENCE LANDSCAPE, LAYER BY LAYER

Layer 1 (Domain Knowledge): Evidence — Strong. The science of how people acquire, store, and retrieve knowledge is well-established. Retrieval practice, spaced repetition, interleaving, and cognitive load management are evidence-based strategies with large, replicated effect sizes. A curriculum designer can build on this with high confidence.

Layer 2 (Skill): Evidence — Strong to Moderate. Deliberate practice with feedback is the established mechanism for skill development, though its explanatory power varies across domains. The expertise reversal effect (what helps novices may hinder experts) requires adaptive instruction. A curriculum designer can build on this with moderate-to-high confidence but must be sensitive to the domain-specificity of practice effects and the limits of the deliberate practice framework outside well-structured domains.

Layer 3 (Judgment): Evidence — Moderate. The conditions under which expert judgment develops reliably are specified by the Kahneman-Klein framework (high environmental validity + adequate feedback opportunity). Case-based reasoning, simulation, after-action reviews, and exposure to ambiguity and uncertainty are promising approaches supported by moderate evidence. But judgment development requires time and varied experience that cannot be short-circuited, and judgment is not reliably achievable in low-validity or low-feedback domains. A curriculum designer should focus on creating the environmental conditions for judgment development rather than attempting to teach judgment directly.

Layer 4 (Metacognition): Evidence — Moderate to Strong. Metacognitive training — through self-explanation, calibration exercises, prediction-first pedagogy, and productive failure — has a substantial evidence base. The Dunning-Kruger problem means that metacognitive training is most needed by those least likely to seek it, but productive failure provides a mechanism for making knowledge gaps visible even to those with poor metacognitive monitoring. A curriculum designer should embed metacognitive training in content instruction rather than treating it as a separate curriculum component.

Layer 5 (Character and Disposition): Evidence — Thin. Intellectual humility is measurable and associated with better epistemic outcomes, but intervention studies are almost nonexistent. The strongest evidence points toward environmental design — creating conditions that reward honesty, model humility, tolerate uncertainty, and treat error as information — rather than direct training of character traits. A curriculum designer should prioritize environmental design for this layer and be honest about the limits of what direct instruction can achieve.

Environmental Dimension: Evidence — Strong. The evidence that institutional environments promote or degrade competence at layers 3–5 is robust and convergent across multiple research traditions (psychological safety, error management culture, organizational learning, safety science). Environmental design is a first-order educational intervention, not a secondary consideration. A curriculum designer who ignores the learning environment while perfecting content delivery and instructional methods is optimizing the wrong variable for full-stack competence.

9.2 THE INTEGRATION PROBLEM

The most important challenge for a curriculum designer committed to full-stack competence is integration. The five layers are not independent. They interact in complex, bidirectional ways that a simple "stack" metaphor does not fully capture.

Knowledge enables judgment, but does not guarantee it. You cannot judge what you do not understand — the surgeon needs anatomy, the pilot needs aerodynamics, the teacher needs pedagogy. But knowledge accumulation alone does not produce judgment. The Dreyfus model's stages describe a qualitative transformation that goes beyond "knowing more." The Chi expert-novice paradigm demonstrates that expert knowledge is not merely larger but differently organized. This reorganization — from surface-feature to deep-structure representation — is the mechanism through which knowledge becomes judgment-enabling, and it requires experience with varied problems, not just more information.

Metacognition enables skill development, and skill development enables metacognition. Accurate self-monitoring is required for effective practice — you must know what you are doing wrong in order to correct it. But the Dunning-Kruger findings show that the capacity for accurate self-monitoring develops *with* domain competence, not prior to it. This creates a chicken-and-egg problem: you need metacognition to develop skill effectively, but you need skill to develop metacognition accurately. The practical resolution is that metacognitive training and content training should proceed in parallel, not sequentially, with each supporting the other.

Environment shapes everything, multiplicatively. A punitive environment degrades metacognition (people stop honestly assessing their own performance because honest self-assessment is risky), suppresses character (people learn to perform confidence rather than exercise humility because genuine uncertainty is punished), and impairs judgment development (feedback loops are severed because error reports carry social costs). Conversely, a psychologically safe, error-tolerant environment enables the development of all upper layers simultaneously. The L1-002 finding that controlling environments undermine intrinsic motivation, the L1-003 finding that high-stakes assessment distorts learning, and the L1-009 finding that toxic environments manufacture incompetence at layers 3–5 are not three separate findings — they are three perspectives on a single underlying phenomenon.

Productive failure operates at multiple layers simultaneously. It builds metacognitive awareness (layer 4) by making knowledge gaps experientially visible. It activates and reveals the state of domain knowledge (layer 1). It develops tolerance for uncertainty and frustration (layer 5) through the experience of struggling with genuine difficulty. And when followed by instruction, it contributes to the pattern recognition base that supports judgment (layer 3) because the instruction connects to richly activated prior knowledge. Productive failure is not a technique for training any single layer — it is a design principle that leverages the interactions between layers.

The diagnostic questions interact. COMPETENCE-TARGET.md's five diagnostic questions — Do they know what good looks like? Can they tell when reality differs? Do they care? Will they update? Are they allowed to tell the truth? — are not independent checks. A "no" at any point propagates. If they are not allowed to tell the truth (question 5), they will stop caring about accuracy (question 3), which degrades their ability to tell when reality differs (question 2), which eventually compromises their knowledge of what good looks like (question 1), because the feedback loops that maintain all layers have been severed.

The practical implication is that a curriculum designed for full-stack competence cannot simply stack five separate interventions — one for each layer. It must create learning experiences that develop multiple layers simultaneously and, crucially, must attend to the environmental conditions

that enable or disable the upper layers. The environment is not a backdrop to instruction; it is the medium through which instruction either succeeds or fails at the upper layers.

9.3 WHAT WE STILL DON'T KNOW

Several critical questions remain unanswered:

1. **The training-vs.-environment question for layer 5.** Is epistemic character best understood as a trait that can be trained, a state that is facilitated by environment, or both? The current evidence leans toward the environmental explanation, but this may partly reflect the absence of rigorous training studies rather than evidence that training does not work.

2. **Long-term durability.** Most metacognitive and judgment training studies measure short-term outcomes. Do these effects persist over months and years? Do they transfer to novel domains? The longitudinal evidence is sparse.

3. **Individual differences.** How do personality, prior experience, developmental stage, and cultural background moderate the effectiveness of competence formation interventions at layers 3–5? The current literature treats these moderators superficially.

4. **The assessment gap.** Until layers 3–5 can be assessed with acceptable validity and efficiency, institutions will continue to optimize for layers 1–2 by default. Developing practical assessment instruments for the upper layers is as important as developing interventions.

5. **Cross-cultural validity.** Almost all research cited here was conducted in Western contexts. Whether the findings — particularly regarding intellectual humility, psychological safety, and error management culture — generalize to non-Western educational contexts is an open empirical question.

10

CROSS-REFERENCES: WHAT OTHER L1 AGENTS ESTABLISHED

This investigation sits at the intersection of every other L1 agent in the lab. The cross-cutting connections are worth making explicit, because the coherence of the findings across independent investigations is itself evidential — it suggests that the lab's framework is capturing something real about how learning and competence work.

L1-002 (Motivation and Self-Regulation) established that self-determination theory is the strongest framework for understanding motivation; that controlling environments reliably undermine intrinsic motivation; that self-regulation can be taught through direct instruction; and that motivational decline across schooling years is real and correlates with increasing institutional control. This investigation found that the same controlling environments that undermine motivation also prevent the development of upper-layer competence. The motivational decline is not merely a motivational problem — it is a competence-formation problem. SDT's three basic needs (autonomy, competence, relatedness) map onto the environmental conditions that the competence formation literature identifies as essential for developing judgment, metacognition, and character. The convergence is strong and theoretically coherent.

L1-003 (Assessment and Feedback) established the assessment paradox: assessment is simultaneously the most powerful lever for learning and the most common mechanism for undermining motivation. This investigation extends the paradox upward. The same assessment systems that distort motivation also distort metacognition — when assessment rewards correct answers over honest self-evaluation, learners optimize for appearing knowledgeable rather than for *being* accurately calibrated about what they know. The L1-003 finding that task-level and process-level feedback is more effective than self-level feedback maps directly onto the metacognitive training literature: feedback that helps learners understand *what* they got wrong and *why* develops metacognitive accuracy, while feedback that evaluates the learner as a person undermines the psychological safety needed for honest self-assessment.

L1-008 (What Should Be Learned) established the philosophical foundations for curriculum content selection, drawing on Nussbaum's capabilities approach, Dewey's philosophy of growth, and the Bildung tradition. This investigation is the empirical complement: L1-008 asks what should be learned; L1-009 asks what it means to have actually learned it — what depth of acquisition constitutes genuine competence. The convergence between the philosophical traditions L1-008 reviewed and the empirical findings here is notable. Nussbaum's emphasis on "practical reason" as a central capability maps onto layer 3 (judgment). Dewey's insistence that education should produce "continued capacity for growth" maps onto layers 4–5 (metacognition and disposition). The Bildung tradition's claim that education should form the whole person through encounter with substantive content maps onto the full-stack commitment. The philosophical frameworks and the empirical evidence point in the same direction: education that addresses only knowledge and skill while neglecting the upper layers is insufficient by any serious philosophical or empirical standard.

L0-001v2 (Full-Field Survey) identified several gaps that this investigation directly addresses: the transfer problem (Gap 1), self-regulation development (Gap 7), deliberate practice outside expert performance (Gap 10), and cognitive science in ill-structured domains (Gap 5). This investigation's finding that judgment is domain-specific and environment-dependent — not a generic transferable skill — connects to the transfer gap: the reason far transfer is so elusive may be

precisely that the upper layers of competence are domain-specific and context-sensitive, resisting the kind of abstraction that would make general transfer possible.

11 CLOSING ASSESSMENT: CONFIDENCE LEVELS

This section provides an honest accounting of the confidence level attached to each major finding, following the lab's standard confidence framework.

11.1 HIGH CONFIDENCE FINDINGS

- Retrieval practice, spaced repetition, and deliberate practice with feedback are effective mechanisms for building knowledge and skill (layers 1–2)
- Expert-novice differences in problem representation are well-established and reflect qualitative differences in mental organization, not merely quantitative differences in knowledge
- The Kahneman-Klein conditions for intuitive expertise (environmental validity + feedback opportunity) are the most evidence-based framework for understanding when judgment is reliable
- Metacognitive training improves learning outcomes, with moderate effect sizes ($d \approx 0.3–0.5$ for various interventions)
- The Dunning-Kruger effect is real (though partly inflated by statistical artifacts) and represents a genuine metacognitive problem for competence formation
- Productive failure produces superior conceptual understanding and transfer compared to direct instruction alone, with large effect sizes
- Psychological safety is a robust predictor of team learning behavior, with extensive replication across industries and cultures
- Institutional environments can degrade upper-layer competence by severing feedback loops — evidence from healthcare, aviation, and organizational safety is convergent

11.2 MEDIUM CONFIDENCE FINDINGS

- The Dreyfus model accurately describes the phenomenology of expertise development, though the causal mechanisms remain underspecified
- Case-based reasoning and simulation develop judgment more effectively than lecture-based instruction in professional education domains
- Intellectual humility is a distinct, measurable construct associated with better epistemic outcomes
- Error management culture produces better organizational performance than error prevention culture alone
- The environmental dimension is multiplicative rather than additive — the strongest theoretical claim in COMPETENCE-TARGET.md, supported by convergent evidence from multiple traditions but not directly tested in controlled experimental designs

11.3 LOW CONFIDENCE FINDINGS

- Intellectual humility can be increased through explicit training or educational design (preliminary evidence only, no long-term studies)

- Epistemic character (layer 5) can be cultivated through educational interventions as opposed to environmental design alone
- After-action reviews and premortem techniques improve judgment (face validity and widespread adoption, but minimal rigorous experimental evidence)
- Reflection and journaling improve metacognition (depends heavily on implementation quality — unstructured reflection may be inert or harmful)
- The mechanisms of the skill-to-judgment transition can be identified precisely enough to design accelerating interventions

11.4 WHAT WE DON'T KNOW

- Whether intellectual humility training produces durable changes
- How competence at layers 3–5 develops longitudinally over years
- Whether the competence stack as currently defined captures all relevant dimensions of competence (tacit knowledge, emotional regulation, and social intelligence may deserve independent status)
- How cultural context moderates competence formation at the upper layers
- How to assess layers 3–5 with sufficient validity and efficiency for institutional use
- Whether the "manufacture of incompetence" through toxic environments is reversible, and if so, on what timescale

12

A NOTE ON THE COMPETENCE STACK ITSELF

This investigation was tasked with grounding Applied Pedagogy's normative commitment to full-stack competence in the best available evidence. The evidence substantially supports the framework — particularly the claims that knowledge and skill are necessary but insufficient, that the upper layers are neglected by educational systems, and that the environmental dimension is a first-order determinant of outcomes.

However, several aspects of the framework could be refined based on the evidence:

1. **The boundary between layers 3 and 4 is blurry.** Judgment requires metacognition (you must monitor your own reasoning to judge well), and metacognition requires judgment (you must judge when your confidence is warranted). In practice, these layers may be more tightly coupled than the stack metaphor suggests.

2. **Emotional regulation may deserve independent status.** The current framework subsumes emotional factors under character and disposition, but the literature suggests that the ability to manage anxiety, tolerate frustration, and maintain emotional equilibrium in the face of challenge is a distinct capacity that mediates the functioning of all other layers. A surgeon who knows, can do, judges well, monitors her own cognition, and is intellectually honest — but who panics under pressure — is not fully competent. This capacity is partially captured by "tolerance for uncertainty" at layer 5 but may warrant more explicit treatment.

3. **Social and relational competence may be underrepresented.** The stack focuses primarily on individual cognitive and character capacities. But competence in most real-world domains is exercised in social contexts — through communication, collaboration, persuasion, and relationship management. The sociocultural dimension (Vygotsky, Lave & Wenger) is not well represented in the current framework.

These observations are flagged for the PI's consideration in dispatch-recommendations.md.

BIBLIOGRAPHY

Alfano, M., Iurino, K., Stey, P., Robinson, B., Christen, M., Yu, F., & Lapsley, D. K. (2017). Development and validation of a multi-dimensional measure of intellectual humility. *PLoS ONE*, 12(8), e0182950.

Andrade, H. (2019). A critical review of research on student self-assessment. *Frontiers in Education*, 4, 87.

Argyris, C. (1977). Double loop learning in organizations. *Harvard Business Review*, 55(5), 115–125.

Benner, P. (2004). Using the Dreyfus Model of Skill Acquisition to describe and interpret skill acquisition and clinical judgment in nursing practice and education. *Bulletin of Science, Technology & Society*, 24(3), 188–199.

Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F., & Winne, P. H. (2018). Inducing self-explanation: A meta-analysis. *Educational Psychology Review*, 30, 703–725.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74.

Carpenter, J. M., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. M. (2018). Domain-general enhancements of metacognitive ability through adaptive training. *Journal of Experimental Psychology: General*, 148(1), 51–64.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132(3), 354–380.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121–152.

Church, I. M., & Samuelson, P. L. (2017). *Intellectual Humility: An Introduction to the Philosophy and Science*. Bloomsbury.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1), 87–114.

Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. Free Press.

Dreyfus, S. E. (2004). The five-stage model of adult skill acquisition. *Bulletin of Science, Technology & Society*, 24(3), 177–181.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention of material. *Learning and Instruction*, 22(4), 271–280.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, 12(3), 83–87.

Edmondson, A. C. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, 44(2), 350–383.

Edmondson, A. C. (2019). *The Fearless Organization: Creating Psychological Safety in the Workplace for Learning, Innovation, and Growth*. Wiley.

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review*, 85(5), 395–416.

Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine*, 79(10), S70–S81.

Ericsson, K. A., & Harwell, K. W. (2019). Deliberate practice and proposed limits on the effects of practice on the acquisition of expert performance. *Frontiers in Psychology*, 10, 2396.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100(3), 363–406.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10), 906–911.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112.

Hofstede, G. (2001). *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations* (2nd ed.). Sage.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, 64(6), 515–526.

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

Kapur, M., & Kinzer, C. K. (2008). Productive failure in CSCL groups. *International Journal of Computer-Supported Collaborative Learning*, 3(4), 369–384.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, 319(5865), 966–968.

Klein, G. (1998). *Sources of Power: How People Make Decisions*. MIT Press.

Klein, G. (2007). Performing a project premortem. *Harvard Business Review*, 85(9), 18–19.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121–1134.

Leary, M. R., Diebels, K. J., Davisson, E. K., Isherwood, J. C., Jongman-Sereno, K. P., Raimi, K. T., Deffler, S. A., & Hoyle, R. H. (2017). Cognitive and interpersonal features of intellectual humility. *Personality and Social Psychology Bulletin*, 43(6), 793–813.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). The operational validity of a video-based situational judgment test for medical college admissions. *Journal of Applied Psychology*, 90(3), 446–455.

Macnamara, B. N., Hambrick, D. Z., & Oswald, F. L. (2014). Deliberate practice and performance in music, games, sports, education, and professions: A meta-analysis. *Psychological Science*, 25(8), 1608–1618.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation* (Vol. 26, pp. 125–173). Academic Press.

Nolen-Hoeksema, S. (2000). The role of rumination in depressive disorders and mixed anxiety/depressive symptoms. *Journal of Abnormal Psychology*, 109(3), 504–511.

Nussbaum, M. C. (2011). *Creating Capabilities: The Human Development Approach*. Harvard University Press.

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, 8, 422.

Polanyi, M. (1966). *The Tacit Dimension*. Doubleday.

Porter, T., & Schumann, K. (2018). Intellectual humility and openness to the opposing view. *Self and Identity*, 17(2), 139–162.

Porter, T., Elnakouri, A., Meyers, E. A., Shibayama, T., Jayawickreme, E., & Grossmann, I. (2022). Predictors and consequences of intellectual humility. *Nature Reviews Psychology*, 1, 524–536.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68–78.

Salas, E., Rosen, M. A., & DiazGranados, D. (2010). Expertise-based intuition and decision making in organizations. *Journal of Management*, 36(4), 941–973.

Senge, P. M. (1990). *The Fifth Discipline: The Art and Practice of the Learning Organization*. Doubleday.

Strobel, J., & van Barneveld, A. (2009). When is PBL more effective? A meta-synthesis of meta-analyses comparing PBL to conventional classrooms. *Interdisciplinary Journal of Problem-Based Learning*, 3(1).

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive Load Theory*. Springer.

van Dyck, C., Frese, M., Baer, M., & Sonnentag, S. (2005). Organizational error management culture and its impact on performance: A two-study replication. *Journal of Applied Psychology*, 90(6), 1228–1240.

Vaughan, D. (1996). *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*. University of Chicago Press.

Whitcomb, D., Battaly, H., Baehr, J., & Howard-Snyder, D. (2017). Intellectual humility: Owning our limitations. *Philosophy and Phenomenological Research*, 94(3), 509–539.

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, 10, 3087.