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

W2-009 · April 2026

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

CONTENTS

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 across this lab's 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.

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 review 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 what was, until recently, an almost entirely aspirational domain (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.

Three findings distinguish this v2 review from its v1 predecessor:

First, **the environment is constitutive, not merely contextual.** Across sources — Vygotsky (1978)●, Lave and Wenger (1991)●, Edmondson (1999, 2019)◑, van Dyck et al. (2005)○, Dehaene (2020)◑, Senge (2006)◑, and Reason (1997)○ — the evidence converges that without psychologically safe, error-tolerant, feedback-rich environments, competence at every layer either fails to develop or actively degrades. The environmental dimension is not something that surrounds competence formation; it is the medium through which competence forms.

Second, **Layer 5 has emerged from pure aspiration to early-stage empirical science.** McLoughlin, Thoma, and Kristjánsson (2025)●provide the first large-scale empirical validation of Aristotelian phronesis (practical wisdom) as a measurable, multidimensional psychological construct that predicts flourishing. Darnell et al. (2019)●establish the knowledge-action gap as the central problem for character education: moral knowledge explains only about 10% of the variance in

moral behavior. The centrality of moral identity and moral emotion — not moral perception or moral reasoning — as the organizing center of practical wisdom challenges the cognitive emphasis of both the competence stack and the broader Western philosophical tradition.

Third, **artificial intelligence has introduced a new threat to competence formation.** Fan et al. (2024)°, Stadler et al. (2024)°, and Bauer et al. (2025)°converge on the finding that AI tools can degrade the metacognitive engagement that produces deep learning — a phenomenon Fan et al. call "metacognitive laziness." This is not a technology problem but a design problem: AI that provides *information* produces less cognitive offloading than AI that provides *recommendations*. The distinction maps directly onto Vygotsky's framework: tools that scaffold (temporary support enabling internalization) are fundamentally different from tools that substitute (permanent external supports replacing internal capacity).

The central finding of this review can be stated in advance: we know a great deal about building knowledge and skill, a moderate amount about developing judgment and metacognition, and less — but substantially more than v1 had — about cultivating epistemic character. The evidence across all layers suggests that the environmental dimension, particularly the presence or absence of psychological safety, error tolerance, and valid feedback, may be the single most important factor in whether upper-layer competence develops or degrades. An environment that penalizes honesty does not merely suppress truth-telling; over time it degrades the capacity to perceive truth. This is the deepest finding the competence stack has to offer, and it is now supported by evidence from neuroscience (Dehaene), organizational behavior (Edmondson, van Dyck), developmental psychology (Vygotsky), social learning theory (Lave and Wenger), systems dynamics (Senge), safety science (Reason), and the emerging AI and learning literature (Fan, Stadler, Bauer).

W2-008's curriculum review established what competence should look like — the normative framework of what should be learned. This review investigates the mechanisms by which that competence is formed, the conditions under which it thrives, and the processes by which it degrades. The two reviews are complementary: W2-008 asks "what capabilities matter?"; this review asks "how are those capabilities built?"

2

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

The cognitive science of knowledge acquisition and skill development is the strongest area in educational research. This section is deliberately concise — not because the findings are unimportant, but because they are well-established and well- covered in this lab's earlier reviews (L1-001, L1-004). The findings are mature, well-replicated, and practically useful — the clearest case where educational research has produced genuine, actionable knowledge. The primary contribution of this section is the skill-to-judgment transition — how mental representations reorganize from surface-feature to deep- structure encoding, and why this reorganization is the mechanism by which Layer 2 competence becomes Layer 3 judgment.

2.1 THE COGNITIVE ARCHITECTURE

Thinking operates through the interaction of a severely capacity-limited working memory — approximately four items (Cowan, 2001) — with an expansive long-term memory organized into chunks, schemas, and mental representations. Willingham (2021) establishes that background knowledge is not a supplement to thinking but its prerequisite: the baseball expertise study (Recht and Leslie, 1988) demonstrated that students with baseball knowledge comprehended baseball-themed passages better than non-experts regardless of standardized reading level. "Memory is the residue of thought" — students remember whatever they actually think about during an activity, not what teachers intend them to learn.

Knowledge compounds: students with more prior knowledge retain a higher proportion of new information, causing initial gaps to widen exponentially over time — the Matthew effect applied to knowledge (Willingham, 2021). This compounding has equity implications that cascade across all layers of the stack. The "fourth-grade slump" — the comprehension decline of disadvantaged students once assessments presume background knowledge that privileged peers possess — is a direct consequence.

Automaticity in foundational skills frees cognitive resources for higher-order thinking. Students who count rather than retrieve arithmetic facts exhaust working memory on basics, leaving nothing for problem-solving. This is the cognitive load theory mechanism (Sweller, Ayres, and Kalyuga, 2011) applied to development: you cannot engage in judgment (Layer 3) while your working memory is consumed by the mechanics of skill (Layer 2).

2.2 DELIBERATE PRACTICE AND ITS BOUNDARY CONDITIONS

Ericsson, Krampe, and Tesch-Römer (1993) identified deliberate practice — structured, effortful practice with feedback, targeting specific weaknesses — as the primary mechanism by which skill develops. The distinction between naive practice (unreflective repetition producing automation and plateau), purposeful practice (goal-directed with feedback), and deliberate practice (the full infrastructure of expert instruction, calibrated challenge, and immediate feedback) is essential (Ericsson and Pool, 2016).

The mechanism of deliberate practice is the development of increasingly sophisticated domain-specific mental representations. Expert performance depends on organized cognitive structures encoding meaningful patterns rather than on superior general cognitive capacity. Chess grandmasters hold approximately 50,000 hierarchically organized chunks and vastly outperform novices on realistic positions but not random arrangements (Chase and Simon, 1973, Training-derived; Ericsson and Pool, 2016)◑. Mental representations are domain-specific and do not transfer: Steve Faloon's extraordinary digit memory conferred no advantage for letter sequences.

The brain remains plastic throughout this process. Maguire's longitudinal neuroimaging of London taxi drivers shows posterior hippocampal enlargement only in those completing licensure — strong evidence that deliberate practice physically reshapes brain structures (Ericsson and Pool, 2016)◑. But these adaptations are reversible without maintenance and can carry costs to adjacent capacities.

The popularization of the "10,000-hour rule" through Gladwell's *Outliers* distorted Ericsson's actual findings. Ericsson never claimed a fixed threshold; it was an average for a specific sample of violinists, and the variance was substantial. Macnamara, Hambrick, and Oswald (2014)○ conducted a meta-analysis finding that deliberate practice explained only 26% of variance in games, 21% in music, 18% in sports, 4% in education, and less than 1% in professions. Ericsson and Harwell (2019)○ responded that the meta-analysis used overly broad definitions of "practice" that included activities Ericsson would not classify as genuinely deliberate; when the strict definition is applied, variance explained increases to 29–61% after attenuation correction.

The resolution is likely domain-specific. In domains with clear performance criteria and established training methods — music, chess, sports — deliberate practice explains substantial variance. In domains with ambiguous performance criteria and delayed feedback — professional work, education, policymaking — less variance is explained. This maps directly onto the Kahneman-Klein conditions for intuitive expertise (discussed in Section III): domains with valid feedback structures support deliberate practice; domains without them do not.

A finding of immediate consequence for the competence stack: experience without deliberate practice produces decline, not stagnation. Physicians' quality of care degrades over time without active refinement; traditional didactic continuing medical education has negligible impact (Ericsson and Pool, 2016)◑. Teachers plateau within approximately five years without deliberate practice targeting specific behaviors with external feedback (Willingham, 2021)◑. This is the same phenomenon across professions: experience without deliberate effort does not maintain competence at any layer.

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. Five converging lines of evidence illuminate this transition.

The Dreyfus model proposes five qualitatively distinct stages of development — novice, advanced beginner, competent, proficient, and expert — each characterized by a different mode of perception and decision-making (Dreyfus and Dreyfus, 1986)◑. At the novice stage, the learner follows context-free rules analytically and with detachment. At the competent stage, a critical shift occurs: the performer must choose an organizing plan, which introduces subjective responsibility and emotional investment in outcomes. As Dreyfus describes it: "The competent performer, on the other hand, after wrestling with the question of the choice of a plan, feels responsible for, and thus emotionally involved in, the product of his choice" (§2.3.11)●. This emotional engagement is not

incidental — it is the mechanism by which concrete experience generates the pattern library that underlies later intuitive performance.

At the proficient and expert stages, perception becomes holistic and intuitive. The expert does not decompose situations into features and apply rules; they perceive the situation as a meaningful whole and respond from accumulated experience. Dreyfus calls this "holistic similarity recognition" — "the understanding that effortlessly occurs upon seeing similarities with previous experiences" (§2.3.13)●. The distinction between this and random intuition is crucial: expert intuition "is the product of deep situational involvement and recognition of similarity" (§2.3.14)●, not the reenactment of childhood trauma or irrational conformity.

The Kaplan experiment provides the sharpest demonstration: International Master Julio Kaplan, required to add heard numbers at a rate of one per second (completely jamming his analytical mind), "more than held his own against the master in a series of games" (§2.3.18)●. Expert performance is genuinely non-analytical — when the analytical faculty is completely occupied, expert intuition continues to function.

The expert-novice paradigm provides the cognitive mechanism. Chi, Feltovich, and Glaser (1981)○ demonstrated that experts and novices categorize physics problems using qualitatively different features: experts grouped problems by underlying physics principles (conservation of energy, Newton's second law), while novices grouped them by surface features (inclined planes, springs, pulleys). This is not a quantitative difference (more knowledge) but a qualitative one (differently organized knowledge). The reorganization of knowledge from surface-feature to deep-structural encoding is the cognitive mechanism of the Layer 2→3 transition.

Polanyi's tacit dimension provides the philosophical foundation. "We can know more than we can tell" (1966, §6.0.7)● is not merely a statement about the limits of verbal expression. Polanyi demonstrates that tacit knowledge operates through a "from-to" structure: we attend *from* proximal particulars (subsidiary awareness) *to* their distal meaning (focal awareness). The expert does not *apply* knowledge; they *dwell in* it and attend through it to the situation at hand — what Polanyi calls "indwelling." The probe example illustrates the mechanism: "as we learn to use a probe . . . our awareness of its impact on our hand is transformed into a sense of its point touching the objects we are exploring" (§6.0.24)●. The tool becomes an extension of the knower, and the knower's attention passes through the tool to the world.

This from-to structure operates identically in perception, skill, diagnosis, and scientific discovery: "These two aspects of knowing have a similar structure and neither is ever present without the other" (§6.0.11)●. The implication for education is profound: any attempt to formalize all knowledge to the exclusion of tacit knowing is "self-defeating" because it seeks "the kind of lucidity which destroys its subject matter" (§4.0.10, Sen's Foreword)●. Education systems that try to reduce competence to explicitly stated rules and procedures will systematically undermine the development of Layers 3–5.

Klein's Recognition-Primed Decision model provides the empirical evidence for what tacit judgment looks like in action (discussed in detail in Section III). And **Gallwey's coaching phenomenology** provides the practitioner account: the transition from rule-following to intuitive performance requires "letting go" of explicit control — what Gallwey calls letting Self 2 play without interference from Self 1. "The more you tell someone what to do, the worse they perform" — because explicit instruction activates analytical interference with what has become automatic (Gallwey, 1974)○.

The convergence across these five traditions — phenomenological (Dreyfus), cognitive (Chi), philosophical (Polanyi), empirical (Klein), and practitioner (Gallwey) — is the most robust finding in the competence literature. The skill-to-judgment transition involves a qualitative reorganization of knowledge from surface features to deep structure, accompanied by the progressive transfer of

control from explicit, analytical processing to tacit, intuitive pattern recognition. This transition cannot be directly taught; it must be developed through accumulated experience with varied, consequential situations in environments that provide valid feedback.

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 evidence base begins to thin, not because judgment is poorly understood but because it resists the experimental isolation that produces clean findings. The most relevant literatures are naturalistic decision-making, the conditions for intuitive expertise, and the training of judgment in professional domains.

3.1 HOW EXPERTS ACTUALLY DECIDE

Gary Klein's research program on naturalistic decision-making (NDM) produced the single most important empirical finding for Layer 3: experienced decision-makers in high-stakes situations do not compare options analytically. They use Recognition-Primed Decision making (RPD). Of 156 coded decision points in fireground commanders, recognitional decisions accounted for approximately 80%. Comparative evaluation appeared in only 18 cases, "half concentrated among less experienced commanders at a single incident" (Klein, 1998, §3.1.1–§3.1.7)◑. Independent replications across tank platoon leaders, naval commanders, and airline crews showed RPD dominance ranging from 46% (novice tank platoon leaders) to 96% (naval commanders under severe time pressure) (§7.2.1–§7.2.16)◑.

The mechanism is pattern recognition. Klein's central claim — backed by hundreds of critical decision method interviews — is that intuition is not a paranormal faculty but recognition of patterns from accumulated experience:

> "The commanders' secret was that their experience let them see a situation, even a nonroutine one, as an example of a prototype, so they knew the typical course of action right away. Their experience let them identify a reasonable reaction as the first one they considered, so they did not bother thinking of others. They were not being perverse. They were being skillful. We now call this strategy *recognition-primed decision making*." (§3.1.3–§3.1.4)●

A critical and counterintuitive finding: it is *experts* who use the first option they think of (because their first option is usually good), while *novices* must compare options. "Before we did this study, we believed that novices impulsively jumped at the first option they could think of, whereas experts carefully deliberated about the merits of different courses of action. Now it seemed that it was the experts who could generate a single course of action, while novices needed to compare different approaches" (§3.1.31)●. This inverts the common assumption that expertise means more careful analysis; expertise means faster, more accurate recognition.

When recognition generates a candidate action, evaluation occurs through mental simulation — running the action through in imagination to spot problems. "The lieutenant imagined himself carrying it out. Fireground commanders use the power of mental simulation, running the action through in their minds" (§3.1.30)●. Mental simulation is constrained by working memory to approximately three causal factors and six transition states (§5.0.40–§5.0.48)◑.

The USS Gloucester case provides the most dramatic illustration. Lt Cmdr Michael Riley identified a Silkworm missile from a radar blip indistinguishable from a friendly A-6 aircraft, within seconds, before any objective identification data was available. "Riley confessed that when he first saw the radar blip, 'I believed I had one minute left to live'" (§4.0.31)●. The resolution: Riley detected imperceptible acceleration — A-6s flew at constant speed; the Silkworm accelerated slightly as it came off the coast. "Riley said he felt it was accelerating, almost imperceptibly. That was the clue" (§4.0.40)●. This was below conscious awareness. The knowledge was real and consequential (it prevented the ship's destruction) but resisted articulation.

This connects directly to Polanyi's from-to structure: Riley attended *from* subsidiary awareness of the radar blip's behavior *to* its focal meaning (missile vs. aircraft). He could not articulate what he knew, but he knew it — and he was right.

3.2 THE KAHNEMAN-KLEIN CONDITIONS: WHEN JUDGMENT IS RELIABLE

The most important contribution to the judgment question came from the adversarial collaboration between Daniel Kahneman and Gary Klein — two researchers with fundamentally opposed views on expert intuition. Their 2009 paper, "Conditions for Intuitive Expertise: A Failure to Disagree" (2,333 citations)○, achieved a remarkable convergence.

Kahneman, whose career was built on documenting the systematic errors in human judgment, endorsed Klein's RPD findings for domains that met two conditions. Klein, who had championed expert intuition, conceded that expertise was unreliable in domains that did not meet them. Kahneman's *Thinking, Fast and Slow* (2011)◑ reproduces the core argument:

> "If the environment is sufficiently regular and if the judge has had a chance to learn its regularities, the associative machinery will recognize situations and generate quick and accurate predictions and decisions. You can trust someone's intuitions if these conditions are met." (§29.0.32)●

The two conditions are:

Environmental regularity. The domain must contain stable, learnable patterns — valid cues that repeat with sufficient consistency to be internalized. Chess is a high-validity environment. Firefighting and clinical medicine (in many diagnostic presentations) qualify. Stock markets and long-range political forecasting are low-validity environments where patterns are too noisy and unstable for reliable intuition to develop.

Adequate feedback opportunity. The practitioner must receive clear, timely feedback on the outcomes of their judgments. "Among medical specialties, anesthesiologists benefit from good feedback, because the effects of their actions are likely to be quickly evident. In contrast, radiologists obtain little information about the accuracy of the diagnoses they make" (Kahneman, 2011, §29.0.29–§29.0.30)●. A psychiatrist whose patients drop out and are never followed up operates in a feedback desert.

A critical nuance: expertise is task-specific, not domain- general. "Expertise is not a single skill; it is a collection of skills, and the same professional may be highly expert in some of the tasks in her domain while remaining a novice in others" (§29.0.29)●. A psychotherapist may have genuine expertise in reading immediate patient reactions but no expertise in predicting long-term outcomes.

Where feedback is absent or misleading, experience cultivates not expertise but the *illusion* of expertise. Hogarth's "wicked environments" generate false confidence. Kahneman cites Lewis Thomas's example: "a physician in the early twentieth century who often had intuitions about

patients who were about to develop typhoid. Unfortunately, he tested his hunch by palpating the patient's tongue, without washing his hands between patients. When patient after patient became ill, the physician developed a sense of clinical infallibility" (§29.0.25)●.

The deepest threat to Layer 3 is not the absence of expertise but its illusion. "Subjective confidence in a judgment is not a reasoned evaluation of the probability that this judgment is correct. Confidence is a feeling, which reflects the coherence of the information and the cognitive ease of processing it" (Kahneman, 2011, §27.0.14)●. Kahneman's own experience as a young military interviewer illustrates the point: "We knew as a general fact that our predictions were little better than random guesses, but we continued to feel and act as if each of our specific predictions was valid. I was reminded of the Müller-Lyer illusion . . . I coined a term for our experience: the illusion of validity" (§27.0.10)●. The illusion of validity is a cognitive illusion — knowing it exists does not dispel it.

3.3 TACIT KNOWLEDGE AND THE ARTICULATION PROBLEM

Klein's fireground commanders could not articulate their expertise. Polanyi's philosophical analysis explains why. And a vivid practitioner example illustrates the consequence:

> "We asked what a spongy roof is, and he told us that the heat weakens the supports so the surface feels softer just before it collapses, then drops everyone into the fire below. We asked what a spongy roof feels like, and he answered that he couldn't put it into words. To new firefighters, all roofs feel spongy." (Klein, 1998, §3.0.1)●

The expert perceives distinctions the novice cannot access. The knowledge is real and consequential — it prevents deaths — but resists articulation. Yet Klein's own work demonstrates that tacit expert knowledge is not forever locked inside the expert's head. Cognitive task analysis (CTA) can partially externalize it. The AWACS weapons director workstation redesign — based on CTA with expert operators — produced 15–20% overall performance improvement after only 4.5 hours with the new interface versus 1,500 hours with the original, with 20% fewer hostile strikes and 36% fewer missed missiles (§7.2.33–§7.2.48)◑. This is a remarkable finding: expert pattern recognition, while not fully articulable, can be extracted sufficiently to transform interface design.

Schön's concept of *reflection-in-action* (1983)○ occupies the junction where judgment and metacognition meet. Professional practitioners think in action through a process of on-the-spot experimentation and reframing — different from both technical rationality (applying theory to practice) and retrospective reflection-on-action. When surprise disrupts the expert's knowing-in-action, they reframe the problem and experiment within the action. The surgeon adjusting technique mid-procedure, the teacher pivoting mid- lesson when students' responses reveal a misconception — this is reflection-in-action, the tightly coupled operation of judgment (Layer 3) and metacognition (Layer 4) that v1 identified as a blurry boundary.

The Japanese *kata* tradition offers a non-Western perspective on the same phenomenon. In martial arts, tea ceremony, and craft traditions, mastery through form requires years of repetitive practice of externally prescribed patterns before the practitioner is permitted — or able — to improvise. The *kata* functions as what Vygotsky would call an external mediating sign that, through prolonged practice, becomes internalized as tacit competence. The German *Handlungskompetenz* (action competence) tradition arrives at a similar conclusion from a different starting point: competence is not isolated knowledge or skill but the capacity to act effectively in real situations (Klieme and Leutner, 2006; German)○. The French *compétence* tradition (Le Boterf, 2010)○ insists

that competence is not a property of individuals but of the interaction between person and situation — you are not "competent" in the abstract; you are competent *in a situation*.

3.4 THE DEVELOPMENTAL MECHANISM: FROM SURFACE TO DEEP STRUCTURE

How does a person move from knowledge and skill (Layers 1–2) to judgment (Layer 3)? Five independent research traditions converge on a single mechanism: the qualitative reorganization of knowledge from surface features to deep structural patterns, accompanied by the gradual withdrawal of conscious analytical control.

Chi, Feltovich, and Glaser (1981, 5,207 citations)○ provided the foundational empirical demonstration. In sorting tasks, expert physicists grouped problems by underlying physics principles (Newton's second law, conservation of energy), while novices grouped them by surface features (inclined planes, springs, pulleys). The expert-novice difference is not quantitative — experts do not simply know more facts — but qualitative: they organize knowledge around deep causal structure rather than perceptual similarity. This reorganization determines transfer: experts recognize structural similarities across superficially different problems, while novices are trapped by surface similarity and fail to transfer across problems that look different but share the same deep structure.

Dreyfus and Dreyfus (1986)◑ describe this same reorganization as a progression through five qualitatively distinct stages — novice, advanced beginner, competent, proficient, and expert — each characterized by a different mode of perception and decision-making. At the novice stage, the performer follows context-free rules analytically and without emotional investment. At the competent stage, a critical shift occurs: the performer must choose an organizing plan for the situation, and this choice introduces subjective responsibility and emotional engagement:

> "The competent performer, on the other hand, after wrestling with the question of the choice of a plan, feels responsible for, and thus emotionally involved in, the product of his choice. While he both understands and decides in a detached manner, he finds himself intensely involved in what occurs thereafter. An outcome that is clearly successful is deeply satisfying and leaves a vivid memory of the plan chosen and of the situation as seen from the perspective of the plan. Disasters, likewise, are not easily forgotten." (§2.3.11)●

This emotional engagement is not incidental — it is the mechanism by which concrete experiences generate the pattern library that underlies later intuitive performance. Without the emotional stakes of choosing a plan and owning its consequences, experiences remain abstract and fail to build the recognition capacity that produces judgment. The competent stage is where the learner transitions from following other people's rules to making their own choices and living with the results — and this transition is what produces the experiential base for proficiency.

At the proficient and expert stages, perception becomes intuitive and holistic: "We call the intuitive ability to use patterns without decomposing them into component features 'holistic similarity recognition.' When we speak of intuition or know-how, we are referring to the understanding that effortlessly occurs upon seeing similarities with previous experiences" (§2.3.13)●. Expert judgment is neither rational (based on explicit rules) nor irrational (random or impulsive) — it occupies a third category that Dreyfus calls "arational": "A vast area exists between irrational and rational that might be called arational" (§2.3.21)●. This category is essential for understanding why Layer 3 judgment cannot be reduced to Layer 1 knowledge plus Layer 2 skill.

The Julio Kaplan experiment provides dramatic confirmation. An International Master was required to add spoken numbers at the rate of one per second while simultaneously playing five-

second- a-move chess against a slightly weaker master. "Even with his analytic mind completely jammed by adding numbers, Kaplan more than held his own against the master in a series of games" (§2.3.18)●. When the analytical faculty was completely occupied, expert intuition continued to function at near-full capacity. Expert performance is genuinely non-analytical.

But the analytical trap is real. Stuart Dreyfus's own chess testimony illustrates the danger: as a mathematician, he excelled at analytical chess but resisted fast chess — "it didn't give me time to figure out what to do" (§2.3.10)●. His teammates who played fast chess and studied grandmaster games advanced to proficiency; Dreyfus remained stuck at the competent stage. The implication for education is direct: instructional environments that exclusively reward analytical decomposition may systematically prevent the transition from competence to proficiency, producing what Dreyfus calls "expert novices" — performers who can analyze situations correctly but cannot perceive them intuitively.

Gallwey's *The Inner Game of Tennis* (1974)○ describes this same transition phenomenologically from a coaching perspective. His Self 1 (the conscious, analytical mind) and Self 2 (the body's intuitive capacity) map directly onto Dreyfus's analytical and intuitive modes. Peak performance occurs when Self 2 is allowed to operate without interference from Self 1. The coaching paradox follows: the more you tell someone what to do, the worse they perform, because explicit instruction activates Self 1 interference with Self 2's natural self-correction. Gallwey's method — "notice where the ball hits the racket" rather than "keep your wrist firm" — activates non-judgmental awareness that facilitates tacit learning without disrupting it.

The convergence across these traditions is striking. Chi describes the cognitive reorganization. Dreyfus describes the phenomenological stages. Klein describes the operational result (RPD). Polanyi describes the epistemological structure (from-to). Gallwey describes the experiential quality of the transition. All agree: the Layer 2→3 transition involves releasing conscious control over what accumulated experience has made automatic. It cannot be produced by more instruction, more analysis, or more explicit knowledge. It requires varied, emotionally consequential experience with valid feedback — the conditions Kahneman and Klein identified.

3.5 CAN JUDGMENT BE TAUGHT?

The evidence converges on a nuanced answer: judgment cannot be directly taught, but it can be systematically developed.

The distinction matters. "Teaching" implies transmission — the transfer of knowledge from teacher to learner through instruction. Judgment resists this because it consists of tacit pattern recognition that is destroyed by attempts to make it fully explicit (Polanyi's formalization trap). But judgment *can* be developed through instructional designs that provide concentrated, varied exposure to consequential situations with valid feedback.

Several domains have achieved this:

CRM and aviation training. The U.S. Navy's Top Gun school produced a 12.5-to-1 kill ratio (versus 2-to-1 for the Air Force) through a specific training design: daily simulated dogfights, camera and radar recording, intensive after-action review, and immediate re-engagement (Ericsson and Pool, 2016)◑. This is deliberate practice applied to Layer 3: the simulation provides the varied, consequential situations; the recording provides the feedback; the after- action review makes the tacit decision process partially explicit; and the re-engagement allows immediate application of lessons. Crew Resource Management (CRM) extended this model from individual pilot judgment to team decision- making, with documented improvements in crew coordination and safety across the aviation industry.

Medical case-based reasoning. Medical education's case method — presenting complex clinical scenarios requiring diagnosis, differential reasoning, and treatment decisions — develops diagnostic reasoning more effectively than lecture- based instruction, particularly when cases are varied, discussed collaboratively, and followed by feedback on diagnostic accuracy.

Military after-action reviews. The U.S. Army's AAR process embeds several principles the judgment literature identifies as important: immediate, specific feedback tied to concrete situations; explicit articulation of decision-making processes; psychologically safe error examination; and forced comparison of mental models to reality.

Simulation. 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 task. Senge's "practice fields" — structured environments where teams can experiment and make mistakes without operational consequences — converge with this finding (Senge, 2006).

Cognitive task analysis as bridge. Klein's AWACS redesign demonstrates that while expert judgment cannot be fully articulated, CTA can extract enough of its structure to improve training, interface design, and decision support.

But judgment development has strict boundary conditions. It requires what Kahneman and Klein identified: environmental validity and adequate feedback. In domains where these conditions are absent — long-range forecasting, educational policy, organizational strategy — the appropriate response is not better training in judgment but better decision-making structures: checklists, decision aids, peer review, structured protocols that substitute institutional processes for individual judgment (Gawande, 2009). The surgical safety checklist — which reduced deaths by 47% and complications by 36% in an eight-hospital pilot — demonstrates that institutional design can compensate for individual judgment limitations without increasing individual expertise.

3.6 COUNTER-EVIDENCE AND LIMITATIONS

Klein's RPD research relies on structured retrospective interviews. Actual cognition during high-stakes events may differ from recalled cognition; Klein acknowledges this (§17.3.3–§17.3.5) but has no alternative methodology. The Vincennes case (shooting down Iran Air 655) is discussed but attributed to information design failure rather than expert intuition failure — which raises the question of whether survivorship bias in the sample (interviewing experts who succeeded) inflates the apparent reliability of RPD.

There is no controlled experimental comparison directly pitting RPD-trained versus analytically-trained decision-makers on matched outcomes. The evidence for RPD is descriptive and observational, not experimental. The strongest evidence comes from training programs (Top Gun, CRM) that operationalize RPD principles and show performance improvements, but these lack the controlled designs that would isolate the RPD contribution from confounds like increased practice time, motivation, and selection effects.

The Kahneman-Klein conditions framework is descriptive rather than prescriptive. Knowing that expertise requires environmental regularity and feedback does not tell you *how* to design educational environments that meet these conditions. The bridge from judgment research to instructional design remains incompletely built.

4

LAYER 4: METACOGNITION

Metacognition — awareness of one's own cognitive processes, the ability to monitor performance, recognize the boundaries of knowledge, and update beliefs in response to evidence — is the layer where the competence stack's distinctive contribution becomes clearest. Layers 1–3 describe what competent people *know* and *do*. Layer 4 describes their capacity to *know what they know and don't know*, to monitor their own cognition in real time, and to self-correct when their judgments go wrong.

Dehaene (2020) identifies metacognition as "one of the single most important drivers of academic success." This is a strong claim, but it is supported by a converging evidence base from calibration training, productive failure research, and the Dunning-Kruger literature — even as some of that evidence base is now contested.

4.1 THE NEURAL AND COGNITIVE FOUNDATIONS

Before examining specific metacognitive interventions, it is worth establishing why metacognition is so consequential. Dehaene (2020) identifies four pillars of learning — attention, active engagement, error feedback, and consolidation — that together describe the neural mechanisms through which learning occurs. Metacognition operates across all four: it directs attention (knowing where to focus), maintains active engagement (monitoring comprehension rather than passively receiving), processes error feedback (recognizing when predictions fail), and supports consolidation (knowing when material has been adequately learned and when it needs further rehearsal).

The metacognitive dimension of error feedback deserves special attention. The brain, on Dehaene's account, learns by minimizing prediction errors — it generates hypotheses about the world and adjusts its internal models when those hypotheses are wrong. But this adjustment requires the learner to register the error, which is itself a metacognitive act. If the learner does not notice that their prediction was wrong — because the error was too subtle, because the environment did not provide clear feedback, or because the learner was not monitoring their own predictions — then no learning occurs. Error feedback without metacognitive monitoring is noise; metacognitive monitoring without error feedback is empty.

Ericsson and Pool (2016) provide a complementary perspective from the expertise literature. As performers develop increasingly sophisticated domain-specific mental representations — organized cognitive structures encoding meaningful patterns — those representations become sophisticated enough to enable self-monitoring, error detection, and independent self-correction. This is the Layer 2→4 bridge: the development of skill generates the metacognitive capacity to monitor that skill. Elite performers "develop mental representations sophisticated enough to enable self-monitoring" (agent-brief, Claim 7). A concert pianist who can hear the difference between their own performance and an ideal performance has developed a mental representation that serves simultaneously as a skill structure (Layer 2) and as a metacognitive monitoring device (Layer 4).

This finding has a dark corollary: without ongoing deliberate practice, the quality of self-monitoring degrades. Ericsson documents that physicians' quality of care declines or stagnates over time without deliberate practice — and that traditional didactic continuing medical education has

negligible impact on this decline. Experience alone does not maintain metacognitive calibration; it must be experience with the right structure of feedback and progressive challenge.

4.2 METACOGNITIVE TRAINING: WHAT WORKS

Carpenter, Wilford, Kornell, and Mullaney (2018)[○] provide the most direct evidence that metacognitive ability is trainable and transferable. They demonstrated that metacognitive discrimination — the ability to distinguish correct from incorrect judgments — can be improved through adaptive staircase procedures, and that this training transfers across perceptual domains. Training metacognitive sensitivity in one task improved metacognitive performance in untrained tasks. This is one of the rare demonstrations of genuine cross-domain transfer in cognitive training, and it provides direct evidence that Layer 4 is independently trainable — you can improve metacognition without necessarily improving domain knowledge or skill.

The caveat is important: the training was on perceptual tasks (visual discrimination). Whether metacognitive training transfers to higher-level cognitive domains — reasoning, judgment, professional problem-solving — remains an extrapolation.

The strongest operational evidence for Layer 4 training comes from Tetlock and Gardner's (2015)[○] superforecasting research. In the Good Judgment Project, ordinary people trained in probabilistic reasoning and calibration outperformed professional intelligence analysts with access to classified information. Approximately one hour of training in probabilistic reasoning improved forecasting accuracy by 6–11%. Superforecasters are distinguished not by domain expertise or intelligence but by cognitive style: they update beliefs incrementally, think in probabilities, are actively open-minded, and practice calibration. They treat their beliefs as "perpetual beta" — held firmly enough to act on but loosely enough to revise.

The GJP tournament format provides exactly what Kahneman and Klein identified as necessary for expertise development: regular scoring against outcomes (environmental regularity) with Brier scores providing immediate, unambiguous feedback. Forecasting tournaments create the conditions for genuine Layer 4 development. But the boundary conditions matter: calibration training works best in domains with scoreable outcomes and rapid feedback. Whether it transfers to domains with slow, ambiguous feedback — which is where most of education, medicine, and management operate — is the open question.

4.3 THE DUNNING-KRUGER PROBLEM AND ITS REPLICATION STATUS

Kruger and Dunning's (1999)[○] finding — that people who lack competence in a domain also lack the metacognitive ability to recognize their incompetence — became one of the most cited papers in psychology (6,680 citations, FWCI 51.86). In four studies, bottom-quartile performers estimated themselves at approximately the 62nd percentile. The mechanism: the skills required to produce correct responses are the same skills required to recognize correct responses, so the deficit is self-concealing. Training that improved their skills simultaneously improved their metacognitive calibration.

The Dunning-Kruger effect has been a cornerstone of the competence stack's treatment of Layer 4. But it is now contested. Magnus and Peresetsky (2022)[○] demonstrate that the characteristic DK graph pattern can be explained as a statistical artifact: regression to the mean on bounded scales produces the DK pattern even with random, unbiased self-assessments. When actual performance is low, random self-assessment errors can only go up; when performance is high, they can only

go down. Dunkel, Nedelec, and van der Linden (2022)○ find that after controlling for regression artifacts, the effect is substantially reduced.

The v2 assessment: the Dunning-Kruger effect is contested but directionally supported. The debate is about magnitude and mechanism, not direction — poor performers are disproportionately miscalibrated, even if the degree of miscalibration is smaller than originally reported. The practical implication stands: beginners systematically overestimate their competence, and this overestimation impairs learning because they do not seek help or practice. The self-concealing nature of metacognitive deficits means that Layer 4 interventions cannot rely on self-reflection alone — they require external feedback.

This converges with Dehaene's finding that error feedback is a pillar of learning and with Edmondson's finding that psychological safety enables error acknowledgment. If the learner cannot see their own errors (DK problem), and the environment punishes error disclosure (psychological unsafety), then metacognitive development is doubly blocked: blocked from within by the self-concealing deficit and blocked from without by the fear of exposure.

4.4 PRODUCTIVE FAILURE AS METACOGNITIVE INTERVENTION

Kapur's productive failure framework (2024)◑ provides the most developed instructional design for Layer 4 development. The core paradox: "My research on Productive Failure shows that making learning easy does not always ease learning. If not intentionally designed to leverage failure in the initial stages, learning tends to be shallow and inflexible" (§4.5.8)●.

Productive failure works through four mechanisms — the 4A model:

Activation. Prior knowledge is activated by struggle with a novel problem, even when that knowledge is incomplete or incorrect. This activation creates the hooks to which new knowledge can attach.

Awareness. Failure reveals gaps between what the learner knows and what they need to know. This metacognitive awareness — consciousness of one's own knowledge gaps — is precisely the Layer 4 capacity.

Affect. The emotional engagement from struggle — frustration, curiosity, the experience of "hitting the wall" — strengthens encoding and retention. Neuroscience links this emotional engagement to learning.

Assembly. The consolidation phase where instruction following failure organizes and integrates the activated knowledge. "Note that the power of Productive Failure cannot be realized without assembly. You can activate all the prior knowledge you want, create heightened awareness, build the right affective state, but without a proper assembly, all of it goes to waste" (§12.8.16)●.

The assembly requirement is critical: productive failure without structured consolidation is just failure. The teacher's role in the assembly phase is essential. This maps directly onto Edmondson's learning zone: high standards (the problem is genuinely difficult) combined with high safety (failure is treated as data, not as evidence of deficiency). Kapur and Hattie (2022)○ extend this to instructional design: the "Fail, Flip, Fix, Feed" model — where students first attempt problems, then receive instruction, then receive corrective feedback, then consolidate — outperforms the traditional "flip then active learning" approach.

De Jong et al. (2023)○ provide independent confirmation from thirteen leading scholars: inquiry-based instruction produces better results for conceptual knowledge than direct instruction, but the optimal approach combines inquiry and direct instruction, with guidance personalized to learner characteristics. This consensus converges with the competence stack's insight that different layers

may require different instructional approaches: direct instruction for Layer 1, guided inquiry for Layers 2–3, productive failure for Layer 4.

4.5 WHEN REFLECTION HELPS AND WHEN IT HURTS

Reflection is widely prescribed as the mechanism for metacognitive development. But the evidence is more nuanced than the prescription suggests. Reflection helps when it is structured, specific, and oriented toward future action. It hurts when it becomes performative (reflection-for-compliance rather than reflection-for-understanding), ruminative (a loop of self-critical analysis without resolution), or disconnected from experience (abstract rumination without concrete referent).

Schön's (1983) distinction between reflection-in-action and reflection-on-action is relevant here. Reflection-in-action — contemporaneous monitoring and adjustment during practice — is the distinctively professional metacognitive capacity. Reflection-on-action — retrospective analysis after practice — is useful but risks the rationalization problem Klein identified: experts asked to explain their decisions after the fact typically construct rational-sounding explanations that do not accurately describe their actual cognitive process.

A critical distinction emerges from the productive failure literature: the performance-learning gap. Kapur and Kinzer (2008) demonstrated that groups experiencing initial failure on complex problems outperformed direct-instruction groups on subsequent transfer tasks — but not on immediate performance measures. The failure condition produced better learning despite worse performance. Fan et al.'s (2024) finding replicates this pattern in the AI context: AI-assisted learners produced better essays (higher performance) but showed no advantage in knowledge gain or transfer (no better learning). The pattern is consistent: anything that makes performance easier in the moment — whether explicit instruction before struggle or AI assistance during the task — may degrade the learning that would produce long-term competence. The implication for metacognitive development is direct: interventions should be evaluated on learning outcomes, not performance outcomes, because the conditions that develop metacognition are frequently the conditions that make immediate performance worse.

The most productive reflection occurs in the zone between experience and instruction. Kapur's productive failure places reflection after struggle and before formal instruction — when the learner has an activated awareness of their own gaps and is primed to receive new knowledge. After-action reviews (military) and clinical debriefs (medical) achieve the same structure: reflection embedded in the cycle of action, immediately tied to specific decisions and outcomes, conducted in psychologically safe environments where error examination is the norm.

A further implication: the quality of reflection depends on its temporal position relative to experience. The most effective sequence appears to be experience → structured reflection → instruction → consolidation. When reflection precedes experience (pre-briefs without concrete referent), it tends toward abstraction. When it follows instruction (reflect on what you were told), it tends toward regurgitation. When it is embedded between experience and instruction — the learner has struggled, is aware of gaps, and is about to receive structured knowledge that addresses those gaps — reflection serves its metacognitive function most powerfully.

4.6 SELF-REGULATION REVISED: AN ENVIRONMENTAL PRODUCT

V1 treated self-regulation as a meta-skill that could be "taught through direct instruction." The evidence now requires revision. The marshmallow test — long held as evidence that early self-regulation predicts life outcomes — has been substantially revised. Watts, Duncan, and Quan

(2018) found that the effect is approximately half the original size after controlling for socioeconomic status and family background. Falk, Kosse, and Pinger (2019) conducted a direct comparison of the original findings with the replication.

The implication is significant: delay of gratification at age four predicts later outcomes, but the effect is substantially mediated by family background. Self-regulation is not a stable, context-independent trait but a capacity shaped by environmental conditions — particularly the predictability and warmth of the child's environment. Warm, predictable, low-stress environments develop self-regulation; chaotic, unpredictable environments prevent it. This aligns with Vygotsky's argument that higher psychological functions are socially constituted and with Edmondson's finding that safety enables self-regulation.

W2-008 documented this revision and its implications for curriculum design. For the competence stack, the implication is that Layer 4 (metacognitive self-regulation) is not merely supported by the environmental dimension — it is partly constituted by it. You cannot reliably develop self-regulation in environments that are chaotic, threatening, or unpredictable, regardless of the quality of direct instruction.

4.7 AI AND METACOGNITIVE LAZINESS

The most urgent new finding for Layer 4 comes from the emerging literature on AI-assisted learning. Fan et al. (2024, 296 citations) demonstrated that AI-assisted learners (using ChatGPT) showed different self-regulated learning process patterns compared to learners without AI assistance. The AI group outperformed on essay quality but showed no significant advantage in knowledge gain or transfer. The pattern suggests "metacognitive laziness" — learners offloaded metacognitive processes to the AI tool rather than engaging in them independently.

This finding is not isolated. Stadler, Bannert, and Sailer (2024) found that students using LLMs during scientific inquiry experienced reduced cognitive load but produced lower-quality reasoning. The LLMs reduced the "desirable difficulty" of the task — making the experience feel easier while degrading the learning outcome. This is cognitive load theory in reverse: by removing germane cognitive load (the effortful processing that produces learning), LLMs remove the conditions for deep learning.

Yan et al. (2024, 271 citations, *Nature Human Behaviour*) provide the most comprehensive review of generative AI's promises and challenges for learning, noting that GenAI can provide personalized tutoring, immediate feedback, and adaptive scaffolding — but can also reduce cognitive effort, degrade metacognition, and create dependency. Bauer et al. (2025) extend this analysis, arguing that AI's effects depend heavily on implementation design: AI used as a scaffolding tool (prompting reflection, asking questions) may enhance learning; AI used as a solution tool (providing answers) may degrade it.

The design principle is clear and converges with three independent evidence traditions:

From the automation bias literature, Goddard, Roudsari, and Wyatt's (2011) systematic review found that providing *information* (raw data, decision-relevant features) produces less automation bias than providing *recommendations* (suggested actions).

From Senge's systems thinking (2006), the "shifting the burden" archetype describes exactly this dynamic: symptomatic solutions (AI solves the problem) weaken the system's capacity for fundamental response (the learner's own metacognitive capacity), creating addictive dependency patterns.

From Vygotsky's developmental framework (1978), the critical question is whether AI tools function as *scaffolding* — temporary support that enables internalization, after which "I don't need

pictures anymore. I'll do it myself" (p. 72) — or as *prosthetics* — permanent external supports that replace rather than develop internal capacity. The developmental progression Vygotsky describes — from needing external aids to internalizing the operation — requires that the external support be gradually withdrawn. AI tools that become permanent substitutes for metacognitive engagement prevent this withdrawal and therefore prevent internalization.

The practical implication for the competence stack: AI tools should be designed to provide information, prompts, and questions — not answers. They should function as Socratic partners that activate the learner's own metacognitive processes (Awareness in Kapur's 4A model) rather than as expert systems that bypass them. The distinction between AI-as-scaffold and AI-as-prosthetic may be the most consequential design decision in education for the coming decade.

5

LAYER 5: CHARACTER AND DISPOSITION

Layer 5 — 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 — has the thinnest evidence base in the competence stack. V1 acknowledged this honestly. V2 can report that the evidence base has materially improved, though it remains early-stage compared to the mature science of Layers 1–2.

The improvement comes from two directions: the Aristotelian character education tradition has produced its first serious empirical findings, and the measurement of epistemic virtues like intellectual humility has advanced from philosophical concept to psychometrically grounded construct.

5.1 THE KNOWLEDGE-ACTION GAP: THE CENTRAL PROBLEM

Darnell, Gulliford, Kristjánsson, and Paris (2019)● establish the central problem for Layer 5 with devastating clarity. Blasi's (1980, 1983) reviews found modest correlations between moral reasoning and moral action. Walker (2004, as cited in Darnell et al.) estimates that moral reasoning explains only about 10% of the variance in moral behavior. This is the "gappiness problem" — the gap between knowing what is right and doing what is right.

The knowledge-action gap means that the competence stack cannot assume knowledge (Layer 1) reliably produces character (Layer 5). A person can know the right thing to do and consistently fail to do it — not from weakness of will but from the absence of the mediating psychological structures that connect knowledge to action. These structures, in the Aristotelian framework, constitute *phronesis* (practical wisdom).

Darnell et al. propose four components of phronesis that map with striking precision onto the competence stack:

(i) Moral perception — the ability to perceive what features of a situation are ethically salient and what virtue(s) they call for. This maps onto Klein's RPD model — the morally wise person *sees* the situation differently from the morally ordinary person, just as the expert firefighter sees what the novice cannot.

(ii) Moral adjudication — the ability to weigh competing considerations when virtues conflict. This is the distinctively phronetic capacity and maps onto higher-order judgment.

(iii) Moral identity — a general conception of the good life that furnishes motivational force. Not a grand philosophical system but the ordinary person's grasp of who they aspire to be.

(iv) Emotional regulation — the agent's emotions are calibrated to moral reality. Not emotional suppression but "the infusion of emotion with reason" — seeing the dangerous as fearsome, being pained by injustice.

5.2 THE ARISTOTELIAN DEVELOPMENTAL TRAJECTORY

Kristjánsson (2015)◑ provides the most developed philosophical framework for how character develops. The trajectory runs from non-rational habituation through cognitive refinement to phronesis-guided autonomy:

> "Gradually, however, that initial process, which is basically non-rational and achieved via conditioning, is superseded by a rational process, whereby learners continue to be conditioned, but through a conditioning that is accompanied by description and explanation — leading, over time, to the formation of the learners' own phronesis." (§9.2.3)•

Sherman resolves the apparent paradox — how do you reach the "critical palace" through an "uncritical courtyard?" — by arguing that habituation was never mindless. The rehearsals required for acquiring virtues "must involve the employment of critical capacities, such as attending to a goal, recognizing mistakes and learning from them, understanding instructions, following tips and cues." Thus habituation constitutes "a critical practice: a gradual dynamic process of moral and intellectual sensitisation and integration" (§9.2.5, Sherman quoted)•.

This parallels the Dreyfus model with remarkable precision. Habituation corresponds to the novice and advanced beginner stages (following rules under guidance). Cognitive refinement corresponds to the competent stage (choosing plans, becoming emotionally invested). Phronesis corresponds to the proficient/expert stages (intuitive moral perception from accumulated experience). The same developmental logic applies across the skill-to-judgment and knowledge-to-virtue transitions: rule-following → emotional engagement → internalized pattern recognition.

But Kristjánsson is honest about the limits: "I have yet to find a single book, or even a single journal article, written by an Aristotelian character educator, that gives pride of place to phronesis education" (§9.1.3)•. The philosophical framework is elegant. The operational pedagogy is missing.

5.3 THE FIRST EMPIRICAL FINDINGS

McLoughlin, Thoma, and Kristjánsson (2025)• provide the most significant advance. Using bottom-up factor analysis with nationally representative UK and US samples, they found that phronesis manifests as a 10-factor network — not the theorized 4-factor structure. The ten components are: Moral Deliberation, Moral Integration, Identity Aspirations, Moral Self-Relevance, Emotional Regulation, Negative Moral Emotion, Positive Moral Emotion, Virtue Identification, Situational Moral Relevance, and Situational Moral Irrelevance.

The most consequential finding is what sits at the center of the network. Network analysis revealed that the most central nodes are identity-related (Moral Self-Relevance, Aspired Moral Identity) and emotion-related (Negative Moral Emotion, Positive Moral Emotion), not perception-related or reasoning-related. The peripheral nodes are moral perception variables (Virtue Identification, Situational Moral Relevance).

This has radical implications. If moral identity and moral emotion — not moral reasoning — are the organizing center of practical wisdom, then character education should begin with establishing who one aspires to be and how one feels about moral situations, not with teaching moral reasoning. This is a direct reversal of the Kohlbergian emphasis that has dominated moral education for half a century.

Phronesis predicts flourishing: at a two-month follow-up, phronesis components predicted multiple dimensions of flourishing, adding an average of 13.7% incremental predictive power beyond Moral Foundations Theory. Character Strengths and Meaning/Purpose showed the strongest associations. Phronesis is not merely a philosophical construct — it predicts real-world outcomes.

5.4 PHRONESIS IN PRACTICE: MEDICAL EVIDENCE

The philosophical framework gains concrete grounding from medical practice. Conroy et al. (2021, 43 citations)[○] conducted a major empirical study of phronesis in the decision narratives of physicians, using narrative analysis of doctors' accounts of ethical decision-making. The central finding: doctors feel professionally and personally vulnerable in ethical decision contexts. The exercise of practical wisdom in real clinical situations — deciding when to withdraw treatment, how to balance patient autonomy with medical judgment, when to challenge a colleague's decision — requires not just moral knowledge but moral courage, and that courage depends on institutional support. Physicians in psychologically unsafe environments described avoiding or deferring ethical decisions; physicians in supportive environments described engaging with them more fully. The connection to Edmondson's psychological safety is direct: Layer 5 competence (character in action) requires Layer 6 (environmental) conditions to function.

Kotzee, Paton, and Conroy (2016, 45 citations)[○] raise three theoretical questions about phronesis in medicine that have practical implications. First, is phronesis more akin to thinking or feeling? The McLoughlin et al. (2025) network centrality findings — with moral emotion and moral identity as the most central nodes — suggest it is closer to feeling than to thinking, at least in its organizing structure. Second, can phronesis be communicated, or is it irreducibly individual? The evidence from Klein's cognitive task analysis suggests a middle position: phronesis cannot be fully articulated but can be partially externalized through structured narrative methods. Third, is phronesis needed in all decisions or only ethically fraught ones? The Aristotelian answer is all decisions that involve competing goods — which, in professional practice, is most decisions of consequence.

These medical studies provide something the philosophical literature cannot: evidence of phronesis operating in real professional contexts, under real constraints, with real consequences. They confirm that practical wisdom is not an abstract philosophical construct but a capacity that practitioners exercise — or fail to exercise — in their daily work, and that the conditions under which they exercise it are substantially determined by their institutional environment.

5.5 THE STATE OF THE FIELD: DEFENDING CHARACTER EDUCATION

Kristjánsson, Harrison, and Peterson (2024, 15 citations)[○] update and rebut ten common objections to character education, revisiting a 2013 article. The authors argue that the evidence since 2013 has materially strengthened the case: measurement has improved (the neo-APM from McLoughlin et al. provides the first psychometrically grounded phronesis instrument), the philosophical grounding is more sophisticated (the four-component model replaces vague appeals to "values"), and the empirical base has grown beyond the point where character education can be dismissed as mere aspiration.

They acknowledge new challenges that have emerged: the replication crisis affects character education research as it does all social psychology, measurement remains heavily reliant on self-report instruments that are vulnerable to impression management, and cultural variation in what counts as virtue raises questions about the universality of any character education program. But they argue these challenges do not undermine the fundamental project — they make it more rigorous. The field is maturing from a prescriptive tradition (telling students to be good) toward an empirically informed developmental science (understanding how practical wisdom actually develops and designing interventions that support that development).

This defense matters for the competence stack because it establishes where the field stands: no longer purely aspirational, not yet experimentally validated, but transitioning from philosophical

framework to empirical research program. The honest assessment is that Layer 5 is approximately where cognitive load theory was in the 1980s — the theoretical framework is compelling, the measurement tools are newly available, and the intervention studies that would confirm or refute the framework are only beginning.

5.6 INSTITUTIONAL CHARACTER DEGRADATION

The evidence for how character *degrades* is, ironically, stronger than the evidence for how it develops. Edmondson's (2019) Volkswagen Dieselgate case provides the most fully documented example. VW combined unachievable performance goals with a fear-based hierarchy. Engineers who reported that emissions targets were technically infeasible were punished or sidelined. The result: over forty individuals participated in embedding defeat software in diesel vehicles, producing an estimated fifty-nine deaths and over $185 million in penalties. This is not a story of individually deficient character — it is a story of institutional design that systematically degraded the character of technically competent engineers. The same people who, in a different institutional environment, might have exercised intellectual courage instead learned to suppress it.

Vaughan's (1996) concept of "normalization of deviance" describes the mechanism: when small deviations from standards become routine and are not met with consequences, the deviation becomes the new standard. Over time, what was once recognized as wrong becomes accepted practice. This is character degradation through environmental habituation — the inverse of Kristjánsson's developmental trajectory. Where virtuous habituation builds practical wisdom, deviant habituation destroys it, and the destruction is invisible to those inside the system because their calibration has shifted with the norm.

Reason's (1997) Swiss cheese model completes the picture. Organizational accidents result from the alignment of multiple failures across successive defense layers. When institutions respond to failures by blaming individuals rather than examining systemic conditions, they prevent the learning that would strengthen defenses. The blame cycle is the organizational version of Dehaene's finding that fear-based error feedback destroys neuronal plasticity: institutions that punish error disclosure prevent the metacognitive transparency that maintains competence across all layers.

5.7 INTELLECTUAL HUMILITY: MEASURABLE BUT NOT YET TEACHABLE

Porter et al. (2022, 212 citations, FWCI 40.81) synthesize the evidence on intellectual humility — recognizing that one's beliefs might be wrong. Multiple reliable IH measures now exist, showing convergent validity across self-report, behavioral, and informant-report measures. This is genuine progress — ten years ago, IH was a philosophical concept without psychometric grounding.

Higher IH is associated with better learning (more attentive to information quality, less confirmation bias), better reasoning (more open to counter-evidence, better calibrated confidence), better social outcomes (less interpersonal conflict), and less susceptibility to misinformation.

But the intervention gap is large: despite the evidence that IH is consequential, very few intervention studies demonstrate that IH can be reliably increased through educational or psychological intervention. We can measure it; we cannot yet teach it.

The predictors of IH are suggestive: secure attachment, growth mindset, intrinsic motivation, and exposure to intellectual diversity. These are environmental variables. IH appears to develop when environments support it — converging with Edmondson's psychological safety finding and reinforcing the environmental dimension's primacy.

5.8 MIXED TRAITS: CHARACTER AS IT ACTUALLY EXISTS

Kristjánsson (2015)◑ provides the empirically grounded alternative to all-or-nothing virtue: most people possess "mixed traits" — clusters that, to a smaller or larger degree, resemble the idealized form but incorporate person-specific "enhancers and inhibitors" that influence motivation in trait-relevant ways (§6.2.6)●. Character is not a binary state but a complex of partially developed dispositions with person-specific triggers.

This framing resolves the situationist challenge (Doris, Harman) that behavior is determined primarily by situational variables. Aristotelianism does not predict perfect virtue — it predicts mixed traits that are developable. The situationist evidence does not refute character; it describes the starting conditions that character education must work with.

But self-deception is a formidable obstacle: "we do not have any direct introspective access to the real makeup of our virtue-relevant dispositional clusters. Indeed, research indicates that self-deceptions in this area are common" (Kristjánsson, 2015, §6.2.7)●. Just as Kahneman's illusion of validity means confidence does not track judgment accuracy, self-reported virtue does not track actual character. The Dunning-Kruger problem applies to character: those most deficient in intellectual humility may be least able to recognize the deficiency.

5.9 COUNTER-EVIDENCE AND THE HONEST ASSESSMENT

The counter-evidence at Layer 5 is substantial:

No RCT evidence for durable virtue change. Kristjánsson's book does not cite a single randomized controlled trial showing that any Aristotelian character education program produces lasting changes in virtue traits. The framework remains normative aspiration with emerging empirical support, not validated pedagogy.

McLoughlin et al. (2025) limitations. Online samples (two Western populations only), self-report measures, limited longitudinal window (two months), and all-adult samples. The factor structure may differ in non-Western cultures. The eliminativist challenge — that phronesis is reducible to discrete psychological processes (metacognition plus moral identity) without positing a unitary virtue — remains open. Lapsley (2019)○ argues for cautious collaboration between philosophy and developmental psychology while warning against mistaking philosophical categories for empirical ones.

The phronesis gap in education. Even after the most careful philosophical reconstruction, there is no operational account of how to teach phronesis. This is Layer 5's deepest challenge: we now know what phronesis looks like (the 10-component network), we know it matters (it predicts flourishing), and we know what sits at its center (identity and emotion) — but we do not yet know how to reliably produce it through educational intervention.

The honest assessment: Layer 5 has moved from aspirational to early-stage empirical. The knowledge-action gap is real and large. The phronesis construct is now measurable and predictive. The centrality of identity and emotion suggests that character education should focus on who learners aspire to be and how they feel about moral situations, not on what they know about ethics. But the gap between framework and pedagogy remains wide, and the evidence base, while improved, is still thin compared to the mature science of Layers 1–2.

6 THE ENVIRONMENTAL DIMENSION

The environmental dimension is not merely a context in which competence develops. It is the medium through which competence forms. This is the strongest finding in the v2 review and the one with the deepest implications for institutional design.

V1 treated the environment as a first-order concern — already a departure from reviews that treat it as background. V2 goes further: the evidence now supports the claim that the environment is *constitutive* of competence formation. Without psychologically safe, error-tolerant, feedback-rich environments, competence at every layer either fails to develop or actively degrades. An institution that penalizes honesty does not merely contain incompetent people — over time, it *manufactures* incompetence by severing the feedback loops required for self-correction.

This finding is supported by evidence from six converging traditions: developmental psychology, social learning theory, organizational behavior, neuroscience, systems dynamics, and safety science.

6.1 THE SOCIAL CONSTITUTION OF COMPETENCE

Vygotsky's (1978)• "general genetic law of cultural development" is the foundational claim:

> "Every function in the child's cultural development appears twice: first, on the social level, and later, on the individual level; first, *between* people (interpsychological), and then *inside* the child (intrapsychological). This applies equally to voluntary attention, to logical memory, and to the formation of concepts." (p. 57)

This means that all higher cognitive functions — the capacities that populate Layers 2 through 5 — originate in social interaction. Individual competence is always a reconstruction of social competence. The mechanism is internalization — but internalization is not passive absorption. It involves active reconstruction: external operations are radically restructured as they move inward. The process takes time and passes through qualitatively different stages.

The Zone of Proximal Development — "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers" (p. 86)• — defines the space where learning can productively occur. "The only 'good learning' is that which is in advance of development" (p. 89)•. Learning does not wait for readiness; properly organized instruction *creates* readiness by awakening developmental processes that operate only in the context of social interaction.

Lave and Wenger (1991)• push this further. Learning is not the acquisition of knowledge by individuals but increasing participation in communities of practice. "Legitimate peripheral participation" describes how newcomers move from peripheral to full participation, developing identity, knowledge, and skill as inseparable aspects of the same process. "Learning involves the whole person; it implies not only a relation to specific activities, but a relation to social communities — it implies becoming a full participant, a member, a kind of person" (p. 53)•.

This challenges the competence stack's individualist architecture directly. Lave and Wenger argue that competence is a property of participation in communities, not an individual attribute. The five layers describe something real about what develops, but what develops is always shaped

by the social structure of the community in which the development occurs. The social structure — its access arrangements, transparency, power relations, and developmental cycles — determines what can be learned.

Sfard (1998)○ argued persuasively that the acquisition metaphor (competence as individual cognitive development) and the participation metaphor (competence as social membership) are complementary, not competing. The competence stack needs both: the cognitive mechanisms that explain how individuals develop mental representations (Willingham, Ericsson, Chi), and the social structures that determine what opportunities for development exist (Vygotsky, Lave and Wenger, Edmondson). Neither alone is sufficient.

6.2 PSYCHOLOGICAL SAFETY: THE ORGANIZATIONAL FOUNDATION

Edmondson's (1999, Abstract-verified; 2019)◑ research on psychological safety provides the organizational-level evidence. The founding empirical finding: in hospital units, better teams appeared to make *more* errors. "And then came the eureka moment. What if the better teams had a climate of openness that made it easier to report and discuss error? The good teams, I suddenly thought, don't make more mistakes; they report more" (Edmondson, 2019, §4.3.8)●.

Psychological safety — "the belief that the work environment is safe for interpersonal risk taking" (§4.2.5)● — is the primary predictor of learning behavior in work teams, not team efficacy or team resources (Edmondson, 1999)○. The mechanism: safety reduces the interpersonal risk of learning behaviors — asking questions, seeking feedback, experimenting, discussing errors. Without safety, these behaviors are career-threatening.

Edmondson's learning zone framework crystallizes the design principle: psychological safety and performance standards are independent dimensions. High safety combined with low standards produces comfort but not competence. Low safety combined with high standards produces anxiety — the fear-driven environment where errors are concealed, feedback is suppressed, and competence degrades. High safety combined with high standards produces the learning zone — "where people can collaborate, learn from each other, and get complex, innovative work done" (§4.10.6)●. This is not "being comfortable" — "Psychological safety enables candor and openness and, as such, thrives in an environment of mutual respect" (§4.10.1)●.

A critical finding: psychological safety predicts "learn-how" (collaborative learning behaviors — peer feedback, group problem-solving) but not "learn-what" (independent learning behaviors — literature review, individual study). This maps directly onto the competence stack: Layers 1–2 can develop through individual study. Layers 3–5 require social interaction that depends on psychological safety. You can learn facts and procedures alone; you cannot develop judgment, metacognition, and character without the kind of candid interpersonal engagement that psychological safety enables.

Psychological safety "lives" at the level of the group — not the organization (§4.3.12)●. The same organization can simultaneously support and undermine competence development in different teams, depending on local leadership. "These differences in workplace climate shape behavior in subtle but powerful ways" (§4.2.7)●. This means that the competence of leaders — teachers, managers, mentors — is the rate-limiting factor for everyone else's competence development.

6.3 ERROR MANAGEMENT CULTURE

Van Dyck, Frese, Baer, and Sonnentag (2005)○ provide the organizational-level complement to Edmondson's team-level findings. Error management culture — the shared belief that errors are

natural, can be learned from, and should be communicated openly — is positively related to firm performance. In two studies across small and medium-sized companies in two European countries, error management culture predicted firm profitability and survival, controlling for firm size, age, and industry.

The distinction between error management culture (accepting errors, learning from them) and error prevention culture (trying to avoid all errors) is important. Both may coexist, but error management uniquely predicts performance. This is the organizational-level evidence for Dehaene's "error feedback" pillar: error signals drive learning at both the neural and institutional level — but only if the environment permits error acknowledgment.

Dehaene (2020) provides the neuroscience: "Error feedback must be delivered in a fear-free, emotionally safe way, because punitive, stigmatizing practices (for example harsh grading) destroy confidence and neuronal plasticity and so block real learning." The brain cannot learn when it is afraid. Fear activates the amygdala's threat-detection response, diverting cognitive resources from problem-solving. The cognitive resources needed for Layers 3–4 are consumed by self-protective vigilance.

6.4 HOW TOXIC ENVIRONMENTS MANUFACTURE INCOMPETENCE

The evidence converges on a causal chain by which environments degrade competence:

Step 1: Suppression of honest reporting. Fear of punishment prevents the disclosure of errors, concerns, and dissent (Edmondson, 1999, 2019). Volkswagen combined unachievable performance goals with fear-based hierarchy; engineers embedded defeat software rather than disclose technical infeasibility. Forty-plus individuals were implicated, an estimated 59 deaths resulted, and penalties exceeded $185 million (Edmondson, 2019).

Step 2: Loss of system visibility. Without error reports, leaders lose visibility into actual performance. Decisions are made on the basis of distorted information. Mental models diverge from reality — what Senge (2006) calls mental models operating "below awareness." Detroit automakers visiting Japanese factories failed to recognize just-in-time inventory innovations because their mental models of "legitimate manufacturing" filtered out contradictory observations.

Step 3: Normalization of deviance. Small deviations from standards that produce no immediate adverse consequences become accepted as normal. Vaughan's (1996) analysis of the Challenger disaster documented how the Morton Thiokol engineers' concerns about O-ring erosion were progressively normalized until the anomaly became the baseline.

Step 4: Metacognitive erosion. Sustained operation in an environment where reality is denied degrades the capacity to perceive reality. People who operate for extended periods in systems where truth is penalized do not merely learn to keep quiet — they lose the ability to see clearly. Layer 4 (metacognitive monitoring) atrophies because it has been systematically punished.

Step 5: Character adaptation. The dispositions that Layer 5 describes — intellectual honesty, tolerance for uncertainty, willingness to say "I don't know" — are not merely suppressed in toxic environments; they are actively trained out of people and replaced with the opposite dispositions: performing confidence, penalizing dissent, treating error as attack rather than information.

This causal chain explains how organizations can contain individually competent people and yet produce collectively incompetent outcomes. It also explains why the environmental dimension is multiplicative, not additive: it does not merely subtract from individual competence at a fixed rate. It degrades the very feedback mechanisms by which individuals self-correct, producing a downward spiral in which decreasing system visibility leads to worse decisions, which leads to more errors, which leads to more suppression. Reason (1997) models this as the alignment of

"holes" in multiple defense layers — the Swiss cheese model — where active failures (individual errors at the sharp end) combine with latent conditions (institutional failures at the blunt end) to produce catastrophic outcomes.

6.5 ORGANIZATIONAL LEARNING AND ITS RARITY

Senge's (2006) five disciplines framework — systems thinking, personal mastery, mental models, shared vision, and team learning — describes the organizational conditions for collective competence development. Argyris's (1977) concept of double-loop learning — where organizations not only correct errors but question the assumptions that produced them — describes what is required at the institutional level.

The sobering evidence, however, is that genuine organizational learning is rare. Senge's Beer Game simulation, played "thousands of times across five continents," invariably produces identical boom-and-bust cycles regardless of players' backgrounds — demonstrating that systems generate their own crises through structural dynamics, not individual incompetence (Senge, 2006). Argyris's own research at Harvard found that management teams perform adequately on routine decisions but deteriorate when confronting complex, threatening, or embarrassing issues — precisely the issues most requiring their collective intelligence.

Engeström (2001) extends Vygotsky's framework to organizational learning through the concept of "expansive learning" — learning that transforms practice itself, not just the learner's participation in it. This addresses a limitation of Lave and Wenger's framework: their theory describes how newcomers learn existing practice but not how practice is transformed when it is flawed. Institutions that support only reproductive learning produce practitioners who perpetuate existing practice; institutions that support expansive learning produce practitioners who can transform practice when the evidence demands it.

6.6 THE ENVIRONMENTAL EVIDENCE, WEIGHED

The environmental dimension has the strongest convergence of any finding in this review. Vygotsky's developmental psychology, Lave and Wenger's social learning theory, Edmondson's organizational behavior research, van Dyck's error management studies, Dehaene's neuroscience, Senge's systems dynamics, and Reason's safety science all arrive at the same conclusion from different starting points: the environment is not merely a context in which competence develops; it is the medium through which competence forms.

But the evidence has important limitations. Most organizational behavior research (Edmondson, van Dyck) comes from workplace settings. Transfer to educational contexts — classrooms, training programs, mentoring relationships — requires additional argument. Edmondson gestures toward this but does not develop it. Senge's framework, while conceptually powerful, relies on case studies and simulations without controlled evidence. The "learning organization" and its component disciplines lack measurable definitions that would permit replication or falsification. The convergence with stronger evidence (Edmondson, Dehaene, Vygotsky) is what gives Senge's work weight, not its own empirical base.

Despite these limitations, the practical conclusion is robust: environmental design is a first-order educational intervention. A curriculum that develops all five layers of the competence stack requires an institutional environment that meets the conditions for learning: psychological safety, error tolerance, valid and timely feedback, graduated challenge, and transparent access to full

practice. Without these conditions, instruction in knowledge (Layer 1) and skill (Layer 2) may proceed, but judgment (Layer 3), metacognition (Layer 4), and character (Layer 5) cannot develop.

7

CROSS-LAYER INTERACTIONS AND THE REVISION QUESTION

The competence stack presents five layers as conceptually distinct, and this section addresses three questions that the evidence raises about the architecture of the stack itself: How do the layers interact? Are there missing dimensions? Does the stack need structural revision?

7.1 HOW THE LAYERS INTERACT

The layers are not independent. The evidence reveals several interaction patterns:

Bottom-up dependency. Each layer builds on the ones below it, but not mechanistically. You cannot exercise judgment (Layer 3) without knowledge (Layer 1) and skill (Layer 2) to draw on — Klein's expert firefighters would have nothing to recognize without a pattern library built from extensive domain experience. You cannot monitor your own cognition (Layer 4) without having cognition to monitor — Ericsson's finding that elite performers develop mental representations sophisticated enough to enable self-monitoring shows how skill development generates metacognitive capacity. You cannot exercise intellectual honesty (Layer 5) without the metacognitive capacity to recognize your own errors (Layer 4) — which is why the Dunning-Kruger problem, even in its contested form, describes a genuine obstacle to character development.

The knowledge-action disconnect. Bottom-up dependency does not mean bottom-up causation. Darnell et al.'s (2019)● finding that moral knowledge explains only 10% of moral behavior variance means Layer 1 does not reliably produce Layer 5. The mediating mechanisms — identity, emotion, habituation, community membership — must be explicitly addressed. You cannot assume that teaching people about intellectual humility will make them intellectually humble.

The Layer 3–4 tight coupling. V1 flagged the boundary between judgment and metacognition as "blurry." The v2 evidence confirms this coupling but clarifies its nature. Schön's (1983)○ reflection-in-action describes the junction where judgment and metacognition operate simultaneously: the expert practitioner judges and monitors their judging in the same act. The surgeon adjusting technique mid-procedure is simultaneously exercising judgment (Layer 3) and metacognition (Layer 4). This tight coupling is not a problem with the stack but a real feature of professional competence. The layers are analytically distinct (you can have judgment without metacognition and metacognition without judgment) but functionally coupled in expert performance.

Dreyfus's (1986)◑ treatment of the expert's reflective testing of intuitions captures this: "The master chess player contemplates the differences, looking for a move that keeps all intuitively desirable options open while reducing his sense of uneasiness" (§2.3.22)●. This is not analytical problem-solving; it is metacognitive monitoring of intuitive judgment. The boundary between Layers 3 and 4 exists but is permeable in both directions.

Phronesis as cross-layer integration. Kristjánsson's (2015)◑ phronesis occupies the junction of Layers 3, 4, and 5. Phronesis integrates moral perception (Layer 3), moral adjudication (higher-order judgment), moral identity (Layer 5), and emotional regulation (a capacity v1 flagged as potentially missing). The four-component phronesis model maps onto the upper layers of the stack, suggesting that at the highest levels of competence, the layers do not merely interact — they fuse into an integrated capacity that cannot be decomposed without distortion.

Top-down effects are as real as bottom-up dependencies. Metacognition (Layer 4) improves skill execution (Layer 2): the self-monitoring capacity Ericsson describes — where elite performers develop mental representations sophisticated enough to detect and correct their own errors — means that metacognitive development feeds back to improve skill quality. Character (Layer 5) shapes what judgment prioritizes: the morally courageous practitioner asks different questions and notices different features of a situation than the morally indifferent one. Klein's RPD model describes what experts see; character determines what they look *for*. Intellectual humility (Layer 5) enables metacognitive accuracy (Layer 4): the willingness to be wrong is a precondition for honest self-assessment. Without it, metacognitive monitoring becomes self-congratulation rather than self-correction. The stack is not a one-way escalator; it is a dynamic system where development at any layer can support or constrain development at any other.

Environmental mediation of all layers. The environmental dimension does not merely support individual layer development; it mediates the interactions between layers. Psychological safety enables the error acknowledgment that connects Layers 4 and 5 (metacognitive awareness + intellectual honesty). Feedback quality determines whether Layer 2 experience becomes Layer 3 judgment (Kahneman-Klein conditions). Community membership shapes identity development (Lave and Wenger) and thereby affects Layer 5.

7.2 THE REVISION QUESTION: WHAT IS MISSING?

V1 flagged three potential revisions. The v2 evidence supports all three.

Emotional regulation. McLoughlin et al. (2025)• found Emotional Regulation as one of the ten phronesis components. Darnell et al. (2019)• proposed emotional regulation as the fourth component of phronesis — "the infusion of emotion with reason." Dehaene's fear-free error feedback, Edmondson's psychological safety, Dreyfus's finding that emotional engagement at the competent stage is the mechanism for subsequent intuitive development, and the Aristotelian tradition all converge on emotional regulation as essential to competence development across all layers.

The recommendation: emotional regulation is not a separate layer but a cross-cutting capacity that mediates all layer transitions. It belongs in the stack's architecture as a recognized mediating process — analogous to Darnell et al.'s fourth component of phronesis — rather than as a sixth layer. The evidence does not support treating it as independent of the existing layers; it supports treating it as the affective dimension of each layer's development.

Relational and social competence. W2-008's convergence map identified relational capability as the strongest predictor of adult flourishing across independent evidence lines. Lave and Wenger (1991)• argue that competence is participation in communities, not an individual attribute. Vygotsky (1978)• demonstrates that higher functions are socially constituted. Edmondson shows that psychological safety is fundamentally a relational phenomenon.

The recommendation: the competence stack should explicitly acknowledge that competence is always socially embedded. Relational competence is not a missing layer but a dimension that pervades all layers — knowledge is acquired in social contexts, skill is practiced with others, judgment is calibrated through dialogue, metacognition is developed through feedback from others, and character is formed in communities. The environmental dimension already captures the institutional aspect; what needs strengthening is the recognition that the individual side of the stack is always a social product.

Identity. McLoughlin et al.'s (2025)• finding that moral identity is the most central node in the phronesis network — more central than moral reasoning or moral perception — suggests

that identity development deserves explicit attention. Lave and Wenger's (1991) insistence that learning is "becoming a full participant, a member, a kind of person" puts identity at the center of competence formation. Identity is not a separate layer but the integrating thread that connects the learner's developing competence to their sense of who they are and aspire to be.

7.3 THE STACK ARCHITECTURE, RECONSIDERED

The evidence does not support collapsing or merging layers. The five layers describe genuinely distinct capacities: you can have knowledge without skill, skill without judgment, judgment without metacognition, and all four without character. The layers are analytically useful for diagnosis (COMPETENCE-TARGET.md's five diagnostic questions remain valid) and for instructional design (different layers require different pedagogical approaches).

What the evidence does support is three architectural clarifications:

1. **The environmental dimension is foundational, not parallel.** It should be represented as the ground on which the stack stands, not as a separate column alongside it. Without the environmental conditions, the stack does not develop.

2. **Emotional regulation and identity are cross-cutting mediators.** They should be recognized as processes that enable and connect layer transitions, not as additional layers.

3. **The layers are permeable in both directions.** Top-down effects (metacognition improving skill execution, character shaping what judgment prioritizes) are as real as bottom-up dependencies. The stack is not a one-way escalator; it is a dynamic system with bidirectional interactions.

8

ASSESSING LAYERS 3–5

If judgment, metacognition, and character are as consequential as the evidence suggests, they must be assessable with institutional-grade reliability. Otherwise, the competence stack describes a desirable outcome that institutions cannot measure, incentivize, or verify — which means it will be systematically ignored in favor of the layers that *can* be measured (knowledge and skill).

The honest assessment: we can measure some aspects of the upper layers, but the measurement tools are immature compared to the well-developed assessment technology for Layers 1–2.

8.1 JUDGMENT ASSESSMENT

Situational Judgment Tests (SJTs) present realistic scenarios and ask respondents to select or rank response options. They measure practical judgment without requiring full performance in the target situation. The evidence (reviewed in L1-003 assessment investigation) suggests moderate predictive validity for professional performance, incremental to cognitive ability and personality measures. But SJTs are domain-specific (they must be constructed for each profession), expensive to develop, and vulnerable to coaching effects.

Klein's cognitive task analysis provides an alternative approach: rather than testing judgment through standardized scenarios, CTA extracts the judgment patterns that experts actually use. This can inform both training (by making tacit expert knowledge partially explicit) and assessment (by identifying the patterns that distinguish expert from novice judgment). The AWACS redesign (Klein, 1998, §7.2.33–§7.2.48) demonstrates the practical utility of this approach.

8.2 METACOGNITIVE ASSESSMENT

Calibration measures assess the alignment between confidence and accuracy. Well-calibrated individuals have high confidence when they are right and low confidence when they are wrong. The forecasting tournament format (Tetlock and Gardner, 2015) provides a rigorous measurement framework with Brier scores — but only in domains with verifiable outcomes.

Metacognitive questionnaires (Schraw and Dennison's MAI, Panadero's review of self-regulation instruments) provide self-report measures of metacognitive knowledge and regulation. These face the fundamental problem that self-report measures of metacognition are themselves subject to the metacognitive deficits they attempt to measure — the Dunning-Kruger problem applied to measurement.

Process measures — think-aloud protocols, eye-tracking, learning analytics — can capture metacognitive processes in real time without relying on self-report. Fan et al.'s (2024) process-mining methodology for detecting metacognitive laziness in AI-assisted learners represents the frontier of this approach. But process measures are expensive, context-specific, and difficult to scale.

8.3 CHARACTER ASSESSMENT

Intellectual humility scales (Porter et al., 2022)○ provide psychometrically grounded self-report measures with good convergent and discriminant validity. But self-report measures of virtue face the inflation problem Kristjánsson (2015)◑ documents: "self-deceptions in this area are common — namely discordances between actual virtue clusters and virtue-tracking self-concepts" (§6.2.7)●.

The McLoughlin et al. (2025)● neo-APM (neo-Aristotelian Phronesis Measure) represents the most sophisticated instrument for measuring practical wisdom. The 10-factor structure provides a multidimensional assessment with network analysis revealing which components are most central. But it remains a self-report instrument based on responses to hypothetical scenarios.

Behavioral and performance-based measures — observing actual behavior in ethically challenging situations — are the gold standard but are expensive, context-dependent, and raise ethical concerns about deception. The critical incident technique (adapted from Klein's CTA methodology) offers a naturalistic alternative: asking practitioners to describe situations where they faced genuine moral dilemmas and analyzing the quality of their moral reasoning, perception, and action.

8.4 THE ASSESSMENT GAP

The fundamental problem is that the capacities most resistant to direct instruction are also most resistant to standardized assessment. Layers 1–2 are assessable because they involve knowledge and skills that can be elicited through tests and demonstrations. Layers 3–5 involve capacities that are context-dependent, partially tacit, and subject to self-deception — precisely the properties that make standardized assessment difficult.

The practical implication: institutions that want to develop full-stack competence must invest in assessment methods for Layers 3–5 even though these methods are more expensive, less scalable, and less reliable than traditional knowledge tests. The alternative — assessing only what is easy to assess — is the mechanism by which educational systems systematically underweight the upper layers. What gets measured gets managed; what does not get measured gets ignored.

9 PRACTICAL IMPLICATIONS FOR CURRICULUM DESIGN

What does the evidence mean for a curriculum designer tasked with developing full-stack competence? This section distills the review's findings into practical guidance, organized by the design priority the evidence supports.

9.1 PRIORITY 1: DESIGN THE ENVIRONMENT FIRST

The single most consequential design decision is the institutional environment. Before selecting content, designing instruction, or building assessments, the curriculum designer must ensure the learning environment meets the conditions for full-stack competence development:

Psychological safety. Learners and instructors must feel able to ask questions, admit errors, express uncertainty, and challenge ideas without fear of punishment or embarrassment. This is not "being comfortable" — it is creating the conditions under which genuine learning behaviors are possible (Edmondson, 2019)◑. Without psychological safety, Layers 3–5 cannot develop because the interpersonal risk of learning behaviors suppresses them.

Evidence strength: Strong. Edmondson (1999, 10,062 citations), van Dyck et al. (2005, two-study replication), Dehaene (2020, neuroscience convergence).

Error management culture. Errors must be treated as data, not as evidence of deficiency. Error reports must be welcomed and analyzed, not punished. The institutional response to error should be "what can we learn?" not "who is to blame?" (van Dyck et al., 2005; Reason, 1997, Training-derived)○.

Evidence strength: Strong. Van Dyck et al. (two studies), convergence with Dehaene's error feedback pillar and Kapur's productive failure.

High standards alongside safety. Edmondson's learning zone requires both dimensions: safety without standards produces comfort without competence; standards without safety produce anxiety and concealment. The curriculum must demand genuine effort and rigor while making it safe to struggle and fail.

Evidence strength: Strong. Edmondson's 2x2 framework with convergence from Kapur (productive failure requires both challenge and safety) and Vygotsky (the ZPD requires instruction "in advance of development").

9.2 PRIORITY 2: DESIGN LAYER-APPROPRIATE INSTRUCTION

Different layers require different instructional approaches. This is among the review's most practically important findings.

Layer 1 (Knowledge): Direct instruction, retrieval practice, spaced repetition, interleaving. The well-established cognitive science findings apply. Knowledge-rich curricula are prerequisite because knowledge compounds (Willingham, 2021)◑. Automaticity in foundational knowledge frees working memory for higher-order processing.

Evidence strength: Strong. Multiple meta-analyses, well-replicated across age groups and domains.

Layer 2 (Skill): Deliberate practice with feedback, progressive challenge at the edge of ability, worked-example- to-fading sequences. The practice must be genuinely deliberate — struc-

tured, effortful, targeting specific weaknesses — not naive repetition. The boundary conditions of Macnamara et al. (2014) must be acknowledged: deliberate practice explains more variance in well-structured domains than in ill-structured professional contexts.

Evidence strength: Strong in well-structured domains, moderate in ill-structured domains.

Layer 3 (Judgment): Case-based reasoning, simulation, after-action review, productive failure. Varied exposure to consequential situations with valid feedback. The Kahneman-Klein conditions must be met: the learning environment must contain stable, learnable patterns and provide timely, unambiguous feedback. Where these conditions cannot be met, substitute institutional decision structures (checklists, peer review, structured protocols) for individual judgment.

Evidence strength: Moderate. Klein's RPD model (extensive field research but no controlled experiments), Kahneman-Klein conditions (theoretical convergence), Top Gun/CRM (outcome data but no RCTs), simulation training (growing evidence base).

Layer 4 (Metacognition): Calibration training, productive failure, prediction-first pedagogies, structured reflection tied to specific experiences. Kapur's 4A model (Activation, Awareness, Affect, Assembly) provides the mechanism. External feedback is essential because metacognitive deficits are self-concealing (Dunning-Kruger, even in its contested form). AI tools should be designed to prompt metacognitive engagement, not bypass it.

Evidence strength: Moderate. Carpenter et al. (2018, transfer demonstration), Tetlock (calibration training), Kapur (productive failure), Fan et al. (2024, AI warning). The productive failure evidence is strongest in mathematics; generalization to ill-structured domains is less established.

Layer 5 (Character): Begin with identity and emotion, not knowledge. Establish moral self-relevance — who do you aspire to be? — before teaching moral reasoning. Create communities of practice where virtuous dispositions are modeled, practiced, and expected. Acknowledge that character education is early-stage: we know what phronesis looks like (McLoughlin et al., 2025) and what sits at its center (identity and emotion), but we do not yet have validated pedagogies for producing it.

Evidence strength: Weak to moderate. Kristjánsson (2015, philosophical framework), McLoughlin et al. (2025, first empirical validation), Darnell et al. (2019, theoretical framework). No RCT evidence for durable character change.

9.3 PRIORITY 3: DESIGN FOR THE TRANSITIONS

The most critical instructional challenges are not within layers but between them:

The Layer 2→3 transition (skill to judgment). Provide varied exposure to consequential situations with valid feedback. Resist the temptation to reduce complex situations to simple rules. The analytical trap (Dreyfus, §2.3.10)• threatens learners who are rewarded for explicit analysis rather than for developing intuitive pattern recognition. "Fast chess" and exposure to grandmaster games developed proficiency; analytical study kept Stuart Dreyfus stuck at the competent stage.

The Layer 4 development mechanism. Use productive failure (struggle before instruction), calibration training (predict, then check), and structured reflection (not performative journaling but specific, experience-tied, forward-oriented). The AI design principle applies here: tools that do the thinking for learners prevent metacognitive development.

The Layer 5 formation environment. Character develops through communities, not curricula. Create environments where intellectual honesty is modeled and expected, where admitting "I don't know" is valued rather than penalized, and where identity-as-learner is cultivated through the emotional experience of morally salient situations. The competence stack's Layer 5 is an environmental product at least as much as an instructional one.

9.4 PRIORITY 4: TRAIN THE LEADERS FIRST

Edmondson's (2019) most consequential finding for curriculum design may be organizational rather than instructional: psychological safety "lives" at the group level and is "very much shaped by local leaders" (§4.3.12). The same organization can simultaneously support and undermine competence development in different teams, depending on the leader's behavior. If the teacher, mentor, or manager creates an anxiety zone — high standards without psychological safety — then Layers 3–5 cannot develop regardless of curriculum quality.

This makes leader competence the rate-limiting factor for everyone else's competence development. A curriculum designed for full-stack competence will fail if delivered by instructors who punish questions, ridicule uncertainty, or model intellectual arrogance. The practical implication is that any institution serious about developing Layers 3–5 must invest in developing the same capacities in its leaders first — not as an add-on but as a prerequisite.

Marquet's (2012) "intent-based leadership" model on the USS *Santa Fe* provides a practitioner template: by requiring subordinates to state their intentions ("I intend to submerge the ship") rather than asking for orders, Marquet pushed judgment development down to every level of the organization. The preconditions are instructive: clarity (everyone understands purpose and principles) and competence (everyone has the technical knowledge to make good decisions). Without both, decentralized authority produces chaos. With both, it produces an organization where every member practices judgment daily — deliberate practice for Layer 3 embedded in the institutional structure.

9.5 PRIORITY 5: DESIGN FEEDBACK FOR LAYER-APPROPRIATE LEARNING

The evidence supports different feedback designs for different layers:

Layers 1–2 feedback should be immediate, specific, and corrective. Retrieval practice with feedback, worked examples with fading, and deliberate practice with expert guidance all share this structure: the learner makes an attempt, receives clear information about correctness, and adjusts. The feedback loop is tight and unambiguous.

Layer 3 feedback requires ecologically valid settings. Klein's Kahneman-Klein conditions mandate that the feedback environment contain stable, learnable patterns and provide timely outcome information. After-action reviews, case debriefs, and simulation debriefs provide this: they connect the practitioner's decisions to outcomes in a structured format that makes tacit judgment processes partially explicit. The key design principle: feedback must be tied to real or realistically simulated decisions, not to abstract knowledge assessments.

Layer 4 feedback must come from external sources because metacognitive deficits are self-concealing. Calibration measures (Brier scores, prediction-outcome tracking), peer feedback in psychologically safe environments, and structured self- assessment tools that compare self-evaluation to external criteria all serve this function. The Dunning-Kruger problem, even in its contested form, means that asking learners to assess their own metacognitive accuracy without external reference points is circular.

Layer 5 feedback is the hardest to design. Character assessment through self-report faces the inflation problem Kristjánsson (2015) documents: self-deceptions about one's own virtue are common. Behavioral observation in ethically salient situations, peer and mentor feedback on character-relevant behavior, and narrative methods adapted from Klein's critical decision method offer partial solutions — but all are expensive, context-dependent, and resistant to standardization.

9.6 PRIORITY 6: DESIGN FOR MAINTENANCE, NOT JUST DEVELOPMENT

The evidence that experience without deliberate practice produces decline — in physicians (Ericsson and Pool, 2016), in teachers (Willingham, 2021) — has implications beyond initial education. Competence maintenance requires:

- Ongoing deliberate practice targeting specific weaknesses - Periodic recalibration through external feedback - Error management systems that catch and correct drift - Institutional design that prevents normalization of deviance

The Senge "shifting the burden" archetype applies: any system that provides symptomatic relief (outsourcing judgment to procedures, outsourcing metacognition to AI, outsourcing character to compliance systems) risks atrophying the fundamental capacity it claims to support.

9.7 PRIORITY 7: DESIGN AI TOOLS AS SCAFFOLDS, NOT PROSTHETICS

The AI design principle emerges from the convergence of four evidence traditions:

- **Fan et al. (2024):** AI recommendations produce metacognitive laziness. - **Goddard et al. (2011):** Information produces less automation bias than recommendations. - **Senge (2006):** Symptomatic solutions atrophy fundamental capacity. - **Vygotsky (1978):** Scaffolding should enable internalization, after which the scaffold is withdrawn.

The practical application: AI tools in educational contexts should provide information (data, questions, prompts for reflection), not answers (completed essays, solved problems, ready-made analyses). They should function as Socratic partners that activate the learner's own cognitive and metacognitive processes, not as expert systems that bypass them. And they should be designed for eventual withdrawal — the learner should be able to perform without the tool, not become dependent on it.

10

CLOSING ASSESSMENT

10.1 CONFIDENCE LEVELS BY LAYER

Layers 1–2 (Knowledge and Skill): High confidence. The cognitive science is mature, well-replicated, and practically useful. Retrieval practice, spaced repetition, deliberate practice, cognitive load management, and worked-example-to- fading sequences are evidence-based mechanisms with clear instructional implications. The boundary conditions of deliberate practice (Macnamara's meta-analysis) are acknowledged but do not undermine the core findings. The skill-to-judgment transition mechanism — reorganization from surface-feature to deep-structure encoding — is supported by converging evidence from five independent traditions.

Layer 3 (Judgment): Moderate confidence. Klein's RPD model provides a compelling account of how expert judgment operates, supported by extensive field research and cross-domain replication. The Kahneman-Klein conditions establish clear boundary conditions for when judgment is reliable. But the evidence is primarily descriptive (field observation and retrospective interview, not controlled experiment), and the bridge from judgment research to instructional design remains incompletely built. The strongest evidence for judgment development comes from training programs (Top Gun, CRM, medical case-based reasoning) that operationalize the principles but lack the controlled designs that would isolate their specific contribution.

Layer 4 (Metacognition): Moderate confidence. Metacognitive training transfers across tasks (Carpenter et al., 2018). Calibration training works in domains with scoreable outcomes (Tetlock). Productive failure develops metacognitive awareness (Kapur). The DK effect is contested but directionally supported. AI tools can degrade metacognitive engagement (Fan et al., Stadler et al.) — a finding supported by the older automation bias literature. Self-regulation is partly an environmental product (Watts et al., 2018 revision). The evidence is moderate in depth but broad in convergence.

Layer 5 (Character and Disposition): Low to moderate confidence. The evidence has improved dramatically since v1 — from purely aspirational to early-stage empirical. McLoughlin et al. (2025) provide the first large-scale empirical validation of phronesis. The knowledge-action gap (Darnell et al., 2019) establishes the central problem. The centrality of identity and emotion in the phronesis network challenges the cognitive emphasis of traditional moral education. Intellectual humility is measurable but not yet reliably teachable (Porter et al., 2022). The intervention gap remains large: we can describe what phronesis is, measure it, and predict what it produces, but we cannot yet demonstrate a validated pedagogy for producing it.

Environmental dimension: High confidence in principle, moderate in specifics. The convergence across six research traditions (developmental psychology, social learning, OB, neuroscience, systems dynamics, safety science) is the strongest finding in the review. The specific mechanisms — psychological safety, error management culture, learning zone design — are well-supported. The transfer from organizational contexts to educational contexts is plausible but less directly evidenced.

10.2 WHAT V2 RESOLVED THAT V1 COULD NOT

1. **Layer 5 evidence.** V1 had no primary sources for character and disposition. V2 has Kristjánsson (2015), McLoughlin et al. (2025), Darnell et al. (2019), and the intellectual humility literature. Layer 5 has moved from aspirational to early-stage empirical.

2. **The environmental constitutive claim.** V1 treated the environment as first-order. V2 strengthens this to constitutive — supported by full-text engagement with Vygotsky, Lave and Wenger, Edmondson, and convergence with Dehaene's neuroscience.

3. **The AI metacognitive laziness question.** Not addressed in v1. V2 provides a framework (Fan et al., Stadler et al., Goddard et al., Senge's "shifting the burden") with a clear design principle: information not recommendations, scaffold not prosthetic.

4. **The skill-to-judgment transition mechanism.** V1 described this correctly in outline. V2 grounds it in direct engagement with Klein (Sources of Power), Polanyi (The Tacit Dimension), Dreyfus (Mind Over Machine), and Kahneman (Thinking, Fast and Slow), with verbatim quotations from the primary texts.

5. **The DK replication status.** V1 presented the Dunning-Kruger effect without caveat. V2 presents the counter-evidence (Magnus and Peresetsky, Dunkel et al.) and assesses the finding as "contested but directionally supported."

6. **Self-regulation revised.** V1 treated self-regulation as directly teachable. V2 incorporates the marshmallow test revision (Watts et al., 2018) and treats self-regulation as primarily an environmental product.

7. **The stack revision question.** V1 flagged emotional regulation, the Layer 3–4 boundary, and relational competence as issues. V2 proposes specific recommendations: emotional regulation as cross-cutting mediator, Layers 3 and 4 as analytically distinct but functionally coupled, relational competence as pervading all layers rather than constituting a separate one.

10.3 WHAT REMAINS GENUINELY UNKNOWN

1. **How to teach phronesis.** The most important open question. We now know what phronesis is (the 10-component network), what sits at its center (identity and emotion), and what it predicts (flourishing). We do not know how to reliably produce it through educational intervention. The phronesis gap — identified by Kristjánsson himself — remains the central unsolved problem for Layer 5.

2. **Whether calibration training transfers to slow-feedback domains.** Tetlock's superforecasting demonstrates that metacognitive calibration can be trained in domains with scoreable outcomes. Whether this training transfers to education, medicine, management, and other domains where outcomes are ambiguous and feedback is delayed remains an open empirical question.

3. **The long-term effects of AI on metacognitive development.** Fan et al. (2024) and Stadler et al. (2024) document short-term metacognitive laziness. Whether this effect compounds over years of AI-assisted learning — producing a generation of learners with atrophied metacognitive capacity — or attenuates as learners develop AI literacy is perhaps the most important unanswered question for education in the AI era.

4. **Cross-cultural validity of the competence stack.** The evidence base is overwhelmingly Western, Anglophone, and WEIRD. McLoughlin et al.'s (2025) phronesis measure was validated in the UK and US only. The German *Kompetenz* and French *compétence* traditions arrive at similar conclusions from different starting points, which is encouraging. But whether the five-layer

architecture accurately describes competence formation in collectivist cultures, authoritarian institutional contexts, or oral traditions is untested.

5. **The reversibility of competence degradation.** The causal chain by which toxic environments manufacture incompetence is described in Section VI. But how reversible is this process? Can competence degraded by years of operation in a toxic environment be restored? If so, what interventions are effective, and how long does restoration take? The evidence is silent.

6. **Assessment tools for Layers 3–5 at institutional scale.** We can measure judgment (SJTs, CTA), metacognition (calibration, process measures), and character (IH scales, neo-APM) in research contexts. Whether these measures can be scaled to institutional assessment — with the reliability, validity, and cost-effectiveness that institutions require — is an open engineering problem.

7. **The dose-response relationship for environmental interventions.** The evidence supports the claim that psychological safety, error tolerance, and valid feedback are necessary for upper-layer competence development. But how much of each is "enough"? Is there a threshold effect, or is the relationship continuous? Can environmental interventions compensate for deficits in instructional quality, or are both necessary?

10.4 CLOSING

The competence stack describes a real developmental trajectory. The evidence is strongest at the bottom (knowledge and skill), promising in the middle (judgment and metacognition), and emerging at the top (character and disposition). The environmental dimension — the institutional conditions that enable or prevent competence development — is the most consequential and the most actionable finding.

The central practical insight is simple in statement and difficult in execution: to produce people who can work the problem in front of them — who notice, reason from first principles, update on evidence, stay calm under pressure, and know what they don't know — you must design the institutional environment first, the instruction second, and the assessment third. An environment that penalizes honesty will undermine the best curriculum. An environment that rewards honesty and treats error as information will amplify even a mediocre one.

The competence stack is not a finished theory. It is a working framework that organizes a convergent but incomplete evidence base. The five layers are useful abstractions — useful for diagnosis, for instructional design, and for understanding where educational systems succeed and where they fail. The evidence reviewed here confirms the fundamental insight that knowledge and skill are necessary but radically insufficient for real- world competence, and that the institutional conditions in which people learn and work are not merely a backdrop to competence formation but its constitutive medium.

Applied Pedagogy's mission, seen through this lens, is to build educational environments that develop and reward all five layers of the stack — particularly the upper layers that current systems systematically neglect. The evidence says this is possible, but difficult, and honest about where it runs out.

REFERENCES

Andrade, H. L. (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.

Bauer, E., et al. (2025). Looking beyond the hype: Understanding the effects of AI on learning. *Computers & Education,* 221, 105113.

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.

Le Boterf, G. (2010). *Professionnaliser: Construire des parcours personnalisés de professionnalisation.* 6th ed. Éditions d'Organisation.

Carpenter, S. K., Wilford, M. M., Kornell, N., & Mullaney, K. A. (2018). Domain-general enhancements of metacognitive ability through adaptive training. *Journal of Experimental Psychology: General,* 147(5), 767–776.

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.

Conroy, M., et al. (2021). Using practical wisdom to facilitate ethical decision-making: A major empirical study with doctors. *BMC Medical Ethics,* 22, 82.

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.

Darnell, C., Gulliford, L., Kristjánsson, K., & Paris, P. (2019). Phronesis and the knowledge-action gap in moral psychology and moral education: A new synthesis. *Human Development,* 62, 101–129.

Dehaene, S. (2020). *How We Learn: Why Brains Learn Better Than Any Machine . . . for Now.* Viking.

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

Dunkel, C. S., Nedelec, J. L., & van der Linden, D. (2022). Reevaluating the Dunning-Kruger effect. *Intelligence,* 94, 101674.

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.

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

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

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

Engeström, Y. (2001). Expansive learning at work: Toward an activity-theoretical reconceptualization. *Journal of Education and Work,* 14(1), 133–156.

Eraut, M. (1994). *Developing Professional Knowledge and Competence.* Routledge.

Ericsson, A., & Pool, R. (2016). *Peak: Secrets from the New Science of Expertise.* Houghton Mifflin Harcourt.

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.

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.

Falk, A., Kosse, F., & Pinger, P. (2019). Re-revisiting the marshmallow test: A direct comparison of studies by Shoda, Mischel, and Peake (1990) and Watts, Duncan, and Quan (2018). *Journal of Economic Behavior and Organization,* 171, 204–214.

Fan, Y., et al. (2024). Beware of metacognitive laziness: Effects of generative AI on learning motivation, processes, and performance. *British Journal of Educational Technology.*

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

Fleischer, J., et al. (2013). Kompetenzmodellierung: Struktur, Konzepte und Forschungszugänge des DFG-Schwerpunktprogramms. *Zeitschrift für Erziehungswissenschaft,* 16(S1), 5–22.

Gallwey, W. T. (1974). *The Inner Game of Tennis.* Random House.

Gawande, A. (2009). *The Checklist Manifesto: How to Get Things Right.* Metropolitan Books.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2011). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association,* 19(1), 121–127.

Järvelä, S., et al. (2025). Hybrid intelligence: Human-AI coevolution and learning. *Nature Human Behaviour.*

de Jong, T., et al. (2023). Let's talk evidence — The case for combining inquiry-based and direct instruction. *Educational Research Review,* 39, 100536.

Kahneman, D. (2011). *Thinking, Fast and Slow.* Farrar, Straus and Giroux.

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.* Hachette.

Kapur, M., & Kinzer, C. K. (2008). Productive failure in CSCL groups. *International Journal of Computer-Supported Collaborative Learning,* 4(1), 21–46.

Kapur, M., & Hattie, J. (2022). Fail, flip, fix, and feed — Rethinking flipped learning: A review of meta-analyses and a subsequent meta-analysis. *Frontiers in Education,* 7, 956416.

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

Klieme, E., & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. *Zeitschrift für Pädagogik,* 52(6), 876–903.

Kotzee, B., Paton, A., & Conroy, M. (2016). Towards an empirically informed account of phronesis in medicine. *Perspectives in Biology and Medicine,* 59(3), 337–350.

Kristjánsson, K. (2015). *Aristotelian Character Education.* Routledge.

Kristjánsson, K., Harrison, T., & Peterson, A. (2024). Reconsidering the "ten myths" about character education. *Journal of Moral Education,* 53(3), 311–333.

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.

Lapsley, D. (2019). Phronesis, virtues and the developmental science of character. *Human Development,* 62, 101–129.

Lave, J., & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation.* Cambridge University Press.

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.

Magnus, J. R., & Peresetsky, A. A. (2022). A statistical explanation of the Dunning-Kruger effect. *Frontiers in Psychology,* 13, 840180.

Marquet, L. D. (2012). *Turn the Ship Around! A True Story of Turning Followers into Leaders.* Portfolio/Penguin.

Matuschak, A., & Nielsen, M. (2019). How can we develop transformative tools for thought? *Numinous Productions.*

McLoughlin, C. S., Thoma, S. J., & Kristjánsson, K. (2025). Was Aristotle right about moral decision-making? A bottom-up empirical investigation. *PLoS ONE,* 20(1), e0316577.

Nonaka, I., & Takeuchi, H. (1995). *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation.* Oxford University Press.

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

Parasuraman, R., & Manzey, D. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors,* 52(3), 381–410.

Polanyi, M. (1966). *The Tacit Dimension.* University of Chicago Press.

Porter, T., et al. (2022). Predictors and consequences of intellectual humility. *Nature Reviews Psychology,* 1, 524–536.

Reason, J. (1997). *Managing the Risks of Organizational Accidents.* Ashgate.

Schön, D. A. (1983). *The Reflective Practitioner: How Professionals Think in Action.* Basic Books.

Senge, P. M. (2006). *The Fifth Discipline: The Art and Practice of the Learning Organization.* Rev. ed. Currency/Doubleday.

Stadler, M., Bannert, M., & Sailer, M. (2024). Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior,* 160, 108386.

Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction.* Crown.

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

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes.* Harvard University Press.

Watts, T. W., Duncan, G. J., & Quan, H. (2018). Revisiting the marshmallow test: A conceptual replication investigating links between early delay of gratification and later outcomes. *Psychological Science,* 29(7), 1159–1177.

Willingham, D. T. (2021). *Why Don't Students Like School?* 2nd ed. Jossey-Bass.

Yan, L., et al. (2024). Promises and challenges of generative AI for human learning. *Nature Human Behaviour,* 8, 1839–1850.