WHEN THE TRAINER PAYS FOR FAILURE

What Training Science Knows About Learning
That Education Has Missed

Applied Pedagogy Research Lab

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

L1-010 · March 2026

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

CONTENTS

Part I

THE STRUCTURAL ARGUMENT

1

THE INVISIBLE EXPERIMENT

There exists a massive, decades-long natural experiment about how humans learn, and the education research community has largely ignored it.

Every year, the United States military trains hundreds of thousands of people to perform complex, high-stakes tasks — from operating nuclear reactors to flying aircraft in formation to coordinating medical evacuations under fire. Airlines train pilots to manage emergencies that would kill hundreds of people if handled poorly. Hospitals train surgeons to perform procedures where a millimeter of error means a dead patient. Nuclear power plants train operators to manage systems where a wrong decision can render a city uninhabitable.

These training contexts share a structural feature that fundamentally distinguishes them from formal education: **the organization bears the cost of failure, not the learner.** If a military trainee cannot perform, the mission fails and people die. If a pilot trainee cannot handle an engine failure, the airline loses a $200 million aircraft and 200 passengers. If a surgical resident cannot control a hemorrhage, the hospital faces a malpractice suit and a dead patient.

This single structural inversion — who pays when learning fails — changes everything downstream about how learning is designed. It changes the purpose of assessment, the flexibility of time, the speed of feedback, the tolerance for error during practice, and the seriousness with which the institution approaches the question of what actually works. Training organizations are, in a very real sense, forced to figure out what works because they cannot afford to get it wrong.

Education researchers have systematically underweighted this evidence because it lives in different journals (*Human Factors*, *Military Psychology*, *Academic Medicine*, *Journal of Workplace Learning*), different conferences (I/ITSEC, HFES, AMEE), and different academic departments (industrial-organizational psychology, human factors engineering, health professions education). The education-training distinction, which the academy treats as a meaningful categorical boundary, has functioned as an epistemological barrier — preventing the cross-pollination of two traditions that study the same fundamental phenomenon: how humans develop competence.

This dissertation examines what the training tradition knows about learning, evaluates whether that knowledge transfers to educational contexts, and identifies specific design principles that Applied Pedagogy could adopt. It addresses five guiding questions: What does military and aviation training research show about reliably producing competence? What design principles emerge from mastery-oriented training? How do high-reliability organizations train for judgment? What does the workplace learning literature reveal about competence development through practice? And what happens when training principles are imported into educational contexts?

The evidence base is substantial but scattered — spread across military journals, medical education journals, human factors journals, management journals, and organizational psychology journals. The findings are often more rigorous than their educational counterparts — precisely because the organizations producing them had real skin in the game. And the convergence between these findings and those from cognitive science, instructional design, and motivation research provides some of the strongest evidence in the entire learning sciences literature for specific design principles that actually work.

2 THE STRUCTURAL COMPARISON: HOW INCENTIVE INVERSION CHANGES EVERYTHING

Before examining specific training domains, it is worth mapping the structural differences between training and education systematically. These are not differences of degree but of kind, and they cascade through every design decision.

2.1 WHO BEARS THE COST OF FAILURE

In school, if a student fails an exam, the student suffers — a bad grade, a failed course, perhaps delayed graduation. The institution is largely insulated. Indeed, the institution's grading curve may require that some students fail; a course in which everyone succeeds might be suspected of insufficient rigor. The student is the one with skin in the game.

In training, the equation is reversed. If a flight student cannot land an aircraft safely, the airline has invested hundreds of thousands of dollars in training costs and has nothing to show for it. If a surgical resident cannot perform a procedure independently, the hospital must provide additional supervision at considerable cost. If a military unit is not combat-ready, the consequences are measured in lives. The *institution* has skin in the game.

This inversion creates a fundamentally different set of incentives. Training organizations are incentivized to actually figure out what works — not to publish about it, not to win grants about it, but to produce competent performers as efficiently as possible. When the U.S. Army Research Institute studies training effectiveness, it does so because the Army needs soldiers who can do their jobs. When an airline develops CRM training, it does so because the alternative is crashed aircraft. This is not an idealistic commitment to evidence-based practice; it is survival.

2.2 TIME-TO-COMPETENCE VS. TIME-TO-GRADE

Perhaps the most consequential structural difference is the treatment of time. In education, time is the constant and learning is the variable. A semester is 15 weeks regardless of whether students have mastered the material. At the end of those 15 weeks, students receive grades — an evaluative snapshot — and the system moves on.

In training, the relationship is inverted: learning is the constant and time is the variable. A pilot trainee continues training until they can fly the aircraft safely, however long that takes. A surgical resident continues supervised practice until they can perform procedures independently. A military recruit continues basic training until they meet performance standards.

This is not a minor procedural difference. It is the difference between a system that produces a grade distribution and a system that produces competence. Benjamin Bloom recognized this in his seminal "Learning for Mastery" paper (Bloom, 1968), arguing that most students could master most material if given sufficient time and appropriate instruction. The mastery learning literature — reviewed in a meta-analysis by Kulik, Kulik, and Bangert-Drowns (1990) — consistently shows effect sizes of 0.5 to 1.0 standard deviations when time constraints are relaxed and criterion-referenced approaches are used. Training contexts implement mastery learning as a matter of course, not because they have read Bloom, but because releasing an incompetent pilot or surgeon is not an option.

2.3 THE FUNCTION OF ASSESSMENT

In education, assessment serves primarily evaluative and sorting functions. Grades differentiate students, determine who advances, and signal quality to external audiences (employers, graduate schools). Assessment is *about* the student.

In training, assessment serves primarily diagnostic functions. A check ride tells the instructor what the pilot can and cannot do, so that subsequent training can address the gaps. A simulation evaluation tells the surgeon what aspects of the procedure need more practice. Assessment is *about* the training process* — it provides information to improve the system, not to rank the individuals within it.

This distinction maps directly onto the criterion-referenced versus norm-referenced debate that L1-003 investigated. Training assessment is inherently criterion-referenced: there is a performance standard (land the aircraft safely, control the hemorrhage, complete the mission), and the question is whether the trainee meets it. Educational assessment is often norm-referenced: the question is how the student compares to other students, not whether they have achieved any absolute standard.

The implications are profound. Criterion-referenced assessment creates a collaborative dynamic between trainer and trainee — both are working toward the same goal (the trainee meeting the standard). Norm-referenced assessment creates an adversarial dynamic — the assessor's job is to differentiate, and the student's job is to perform. This connects directly to L1-002's finding that controlling assessment undermines intrinsic motivation. Training assessment, because it is diagnostic rather than evaluative, avoids this motivational trap.

2.4 FEEDBACK LOOP SPEED AND CONSEQUENTIALITY

Training environments are characterized by tight, fast, consequential feedback loops. A pilot in a simulator receives immediate feedback — the aircraft responds to control inputs in real time. A surgical trainee receives feedback within seconds — the tissue behaves in ways that reveal the quality of the technique. An after-action review (AAR) happens within hours of the training event, while the experience is fresh.

Educational feedback, by contrast, is typically slow and inconsequential. A student submits an essay and receives feedback days or weeks later. By the time the feedback arrives, the student has moved on to new material, and the feedback is disconnected from the experience of producing the work. L1-003 documented this extensively: feedback effectiveness degrades sharply with delay, and feedback that cannot be acted upon is largely wasted.

The temporal structure of feedback in training contexts is worth examining in detail. In a flight simulator, the feedback loop is measured in seconds — the pilot pulls back on the yoke, and the aircraft responds immediately. This real-time feedback allows for rapid iteration: try, observe the result, adjust, try again. In a check ride, the feedback loop is measured in hours — the evaluator debriefs the pilot immediately after the evaluation. In an after-action review, the loop is measured in hours to days. Even at the longest timescales, training feedback is dramatically faster than educational feedback, and it is always connected to the specific performance that generated it.

Compare this to the typical educational feedback cycle. A student writes a paper over several days, submits it, waits one to three weeks for feedback, receives comments that may or may not be connected to specific moments in their writing process, and is then expected to internalize those comments and apply them to a future assignment on a different topic. The cognitive science of feedback (Hattie & Timperley, 2007, as reviewed in L1-003) makes clear why this process is ineffective: the feedback is too delayed, too disconnected from the production process, and often too generic to support specific improvement. The training world's approach to feedback — fast, specific, embedded in practice, and always followed by an opportunity to try again — is not just a different style. It is a fundamentally more effective architecture for learning.

The consequentiality dimension is equally important. In training, feedback is inherently consequential because it is connected to real performance outcomes. In a flight simulator, failing to respond to an engine fire is immediately and viscerally consequential — the simulator responds with all the cues of an actual emergency. In school, failing to respond correctly to a test question has no consequence beyond the grade. The Kahneman-Klein framework for developing intuitive expertise (Kahneman & Klein, 2009) — which L1-009 explored in detail — specifies that judgment develops in high-validity, high-feedback environments. Training contexts routinely provide both. Schools routinely provide neither.

2.5 THE TERMINOLOGY PROBLEM

The education world draws a sharp line between "education" and "training." This distinction carries enormous normative weight. Education is about developing the whole person, fostering critical thinking, building autonomy. Training is about producing specific behavioral outcomes — it is narrower, more instrumental, and implicitly lower-status. A professor educates; a drill sergeant trains. The distinction maps onto a hierarchy of intellectual prestige that has deep roots in the liberal arts tradition.

Is this distinction defensible? The evidence suggests it is not — or at least, not in the way it is typically deployed. The training tradition, at its best, addresses the full competence stack. Crew resource management training develops judgment (layer 3) and metacognition (layer 4). After-action reviews develop honest self-assessment (layer 5). Medical simulation develops the capacity to perform under uncertainty and pressure — precisely the upper layers of the competence stack that COMPETENCE-TARGET.md identifies as most important and most neglected.

There is a deeper point here. The education world's dismissal of training as "mere skill development" reflects an implicit theory about the hierarchy of learning — that understanding is superior to performance, that knowing *why* is more important than knowing *how*, that the abstract is more valuable than the concrete. This hierarchy has deep roots in the Western philosophical tradition, stretching back to Plato's distinction between *episteme* (true knowledge) and *techne* (craft knowledge). But it is an empirical question whether this hierarchy reflects reality, and the training literature provides evidence that it does not — or at least, not in the simple form that the education world assumes.

Consider what happens when a pilot encounters an engine failure at 3,000 feet. The pilot must understand aerodynamics (domain knowledge), execute emergency procedures (skill), decide between available options based on altitude, weather, terrain, and aircraft state (judgment), monitor their own cognitive state for fixation or tunnel vision (metacognition), and maintain the intellectual honesty to accept that the situation may be unrecoverable rather than persisting in a doomed plan (character). All five layers of the competence stack are activated simultaneously, under extreme time pressure, with real consequences. There is nothing "mere" about this performance. It integrates abstract knowledge and concrete skill in a way that the education world's hierarchy fails to capture.

The education-training distinction has functioned less as an analytical category than as an epistemological barrier — one that has prevented education researchers from learning what the training world knows about developing competence at layers 3–5. The irony is that training contexts may do a better job of producing what liberal education claims to value — judgment, metacognition, intellectual honesty — precisely because they have real feedback loops, real consequences, and environments that reward accuracy over the performance of confidence.

Part II

TRAINING TRADITIONS

3 MILITARY AND AVIATION TRAINING: LESSONS FROM THE HIGHEST STAKES

3.1 THE INSTRUCTIONAL SYSTEMS DESIGN TRADITION

The modern science of instructional design has military origins that the education world has largely forgotten. During World War II, the U.S. military faced an unprecedented challenge: training millions of civilians to perform complex military tasks in a matter of weeks or months. The response was systematic — the military developed formal approaches to analyzing tasks, designing instruction, and evaluating outcomes that eventually crystallized into what is now known as Instructional Systems Design (ISD).

The ADDIE model — Analyze, Design, Develop, Implement, Evaluate — emerged from this tradition, though its exact origins are disputed. Molenda (2003) traced the model's history and found that no single document or author created it; rather, it evolved through multiple iterations of military training development frameworks in the 1960s and 1970s, particularly at Florida State University under contract to the U.S. Army. What matters is not the provenance but the philosophy: training is a *system* that can be designed, tested, and improved through systematic analysis and empirical evaluation.

Robert Gagné's *The Conditions of Learning* (1965, with subsequent editions through 1985) formalized the theoretical foundation for ISD. Gagné identified different types of learning outcomes (verbal information, intellectual skills, cognitive strategies, motor skills, attitudes) and specified different instructional conditions required for each. This taxonomic approach — the idea that different learning goals require different instructional strategies — anticipated the expertise-adaptive model that L1-004 identified as the current evidence-based ideal.

The military ISD tradition produced several insights that education has been slow to adopt:

Task analysis as the foundation of design. Military ISD begins with detailed analysis of what the learner needs to be able to do — not what content they need to be exposed to, but what performance they need to produce. This performance orientation is fundamentally different from the content-coverage orientation that dominates much of education. When the analysis is "the pilot must be able to recover from a stall at 3,000 feet," the instructional design flows directly from that requirement. When the analysis is "students should understand thermodynamics," the instructional design is unconstrained and often unfocused.

Systematic evaluation built into the process. The "Evaluate" step in ADDIE is not an afterthought; it is integral to the system. Military training programs are evaluated against their stated objectives, and programs that do not produce the desired performance are revised. The military uses Kirkpatrick's four-level evaluation framework — reactions (did trainees find the training useful?), learning (did they acquire the target knowledge and skills?), behavior (did they apply what they learned on the job?), and results (did the training improve organizational outcomes?) — as a standard framework for assessing training effectiveness. While few training programs evaluate rigorously at all four levels, the framework itself represents a commitment to accountability that is rare in formal education. Most college courses, by contrast, are evaluated solely through student satisfaction surveys (Kirkpatrick Level 1) — the weakest and least informative level of evaluation. Courses may persist for decades without any assessment of whether students actually learned anything (Level 2), let alone whether they can apply it (Level 3) or whether it contributed to any meaningful outcome (Level 4).

The distinction between training and education within the military itself. It is worth noting that the military draws its own distinction between training (producing specific competencies) and education (developing broader judgment and adaptability). The U.S. military's professional military education system — war colleges, staff colleges, the National Defense University — explicitly aims at the upper layers of the competence stack: strategic judgment, ethical reasoning, leadership under uncertainty. The military has, in effect, a two-track system that addresses the full stack, while education often claims to address the full stack but operationally focuses on layers 1–2.

3.2 CREW RESOURCE MANAGEMENT: THE MOST SUCCESSFUL TRAINING INTERVENTION IN HISTORY

The story of Crew Resource Management (CRM) training is arguably the most compelling case study in the history of learning science — and it is almost entirely unknown in education research.

In the 1970s, the aviation industry faced a crisis. Aircraft were becoming increasingly reliable, but accidents continued to occur at unacceptable rates. Investigation after investigation revealed the same pattern: accidents were caused not by mechanical failure or individual pilot incompetence, but by failures of communication, coordination, decision-making, and leadership within the flight crew. Pilots who were individually excellent were collectively failing.

The aviation psychologist Robert Helmreich and his colleagues at the University of Texas at Austin pioneered the development of CRM training programs (Helmreich, Merritt, & Wilhelm, 1999). CRM training addresses what the aviation community calls "non-technical skills" — communication, situational awareness, decision-making, workload management, teamwork, and leadership. It does not teach pilots to fly; it teaches crews to work together effectively.

Salas, Burke, Bowers, and Wilson (2001) conducted the most comprehensive review of CRM training effectiveness, examining 58 published evaluations using Kirkpatrick's framework (reactions, learning, behavior, results). They found that CRM training consistently produced positive reactions (Level 1), enhanced learning of CRM concepts (Level 2), and promoted desired behavioral changes in the cockpit (Level 3). The evidence for organizational-level outcomes (Level 4 — actual accident reduction) was more difficult to establish rigorously, because accidents are rare events and many variables affect their occurrence. However, the broad trend is unmistakable: the U.S. commercial aviation accident rate declined dramatically during the period when CRM training was widely implemented, from approximately 4 accidents per million departures in the 1970s to fewer than 0.2 per million by the 2010s.

Salas and colleagues updated this review in 2006 (Salas, Wilson, Burke, & Wightman, 2006), finding continued support for CRM's effectiveness while identifying areas needing improvement — particularly in evaluation methodology and in ensuring that CRM principles transfer from the training room to the flight deck.

What makes CRM relevant to Applied Pedagogy is not the specific content of aviation training but the *design principles* it embodies:

Training for the team, not just the individual. CRM recognized that competence is not solely an individual attribute — it is a property of the team system. A crew of individually excellent pilots can still crash an aircraft if they do not communicate, coordinate, and support each other. This insight — that competence has a social dimension — is underappreciated in education, where learning is primarily treated as an individual achievement.

Flattening authority gradients for safety. One of CRM's most important innovations was training subordinate crew members to speak up when they noticed problems, and training captains to listen. The authority gradient in the cockpit — the power differential between the captain and the first officer — was identified as a direct threat to safety. Investigations of multiple accidents revealed that first officers had noticed problems but had been too intimidated to challenge the captain's decisions. CRM explicitly addresses this by training assertion skills and creating norms around mutual monitoring.

This connects directly to Edmondson's (1999) research on psychological safety — the shared belief that a team is safe for interpersonal risk-taking. Edmondson found that teams with higher psychological safety engaged in more learning behavior (asking questions, seeking feedback, experimenting, discussing errors). The FWCI of 42.93 for this paper reflects its enormous impact across multiple fields. CRM can be understood as a systematic intervention to build psychological safety in high-stakes teams.

Scenario-based training with debriefing. CRM training typically uses scenario-based exercises — structured simulations of challenging flight situations — followed by structured debriefing. The scenario provides the experience; the debriefing provides the learning. This two-phase structure (experience followed by structured reflection) is a recurring pattern across effective training programs and resonates with Kolb's experiential learning cycle, Schön's reflective practice, and the after-action review methodology.

Recurrent training, not one-time events. CRM is not a one-time workshop. It is built into recurrent training cycles — pilots receive CRM training and evaluation regularly throughout their careers. This reflects an understanding that non-technical skills degrade without practice and that the organizational culture around communication and teamwork requires ongoing reinforcement.

3.3 AFTER-ACTION REVIEWS: LEARNING FROM EXPERIENCE SYSTEMATICALLY

The After-Action Review (AAR) is one of the most powerful learning tools ever developed, and it originated in the U.S. Army in the 1970s at the National Training Center at Fort Irwin, California. Morrison and Meliza (1999) documented the foundations of the AAR process in their Army Research Institute report.

The AAR is deceptively simple in structure. After a training event (or, increasingly, an actual operation), participants gather to answer four questions:

1. What was supposed to happen? (Plan/intent)
2. What actually happened? (Objective account)
3. Why was there a difference? (Analysis)
4. What will we do differently next time? (Improvement)

The genius of the AAR lies not in these questions but in the cultural norms that surround them. AARs are conducted in a rank-neutral environment — during the AAR, the lieutenant and the general are both just participants analyzing what happened. Honest self-assessment is expected and rewarded. The focus is on processes and decisions, not on blame. The goal is organizational learning, not individual evaluation.

Tannenbaum and Cerasoli (2012) conducted a meta-analysis of debriefing effectiveness across contexts (military, aviation, healthcare, business) and found an overall effect size of approximately $d = 0.67$ — a remarkably large effect for a relatively simple intervention. The meta-analysis also found that debriefing was effective for both teams and individuals, and that facilitated debriefs (with a trained facilitator guiding the discussion) were more effective than unfacilitated ones.

The AAR methodology maps directly onto the competence stack. It develops:

- **Layer 3 (Judgment):** By systematically comparing expected outcomes to actual outcomes, participants develop calibrated mental models of how situations unfold. They learn to recognize patterns, anticipate complications, and adjust their decisions.
- **Layer 4 (Metacognition):** The AAR forces participants to examine their own decision-making processes — not just what they decided, but how and why they decided it. This is metacognitive training embedded in authentic practice.
- **Layer 5 (Character):** The cultural norms of the AAR — rank neutrality, honest self-assessment, focus on learning rather than blame — create an environment that rewards epistemic honesty. Over time, repeated participation in AARs cultivates the disposition to engage honestly with one's own performance.

Why don't schools use anything like the AAR? The answer is structural. Schools lack the performance orientation that makes AARs meaningful — there is no "mission" to debrief, no operational outcome against which to evaluate decisions. Schools also lack the cultural infrastructure that makes AARs work — the norm of rank neutrality, the expectation of honest self-assessment, the focus on collective learning rather than individual evaluation. An AAR in a classroom context would require reimagining the entire relationship between instruction, practice, and assessment.

3.4 SIMULATION AND THE FIDELITY QUESTION

Military and aviation training make extensive use of simulation, from full-motion flight simulators costing tens of millions of dollars to tabletop wargames using paper maps and cardboard counters. The question of how much fidelity (realism) simulation needs to be effective is one of the most practically important questions in the training literature.

The short answer is: less than you think, but more than nothing. The evidence consistently shows that the critical features of simulation effectiveness are not visual or physical fidelity but **functional fidelity** — the degree to which the simulation requires the same cognitive and behavioral processes as the real task. A flight simulator does not need to produce perfect visuals of the terrain; it needs to present the pilot with the same decision-making challenges, the same time pressures, and the same information environment as actual flight.

Gaba (2004) articulated this principle for healthcare simulation, arguing that simulation-based training should be understood not as a technology but as a technique — an approach to experiential learning that can be implemented at many levels of fidelity, from high-tech mannequins to role-playing exercises. What matters is the design of the learning experience, not the sophistication of the equipment.

This insight has profound implications for cost-effective training design and, by extension, for education. If the critical ingredient is cognitive fidelity rather than physical fidelity, then effective simulation-based learning does not require expensive equipment. Case studies, role-playing, structured problem-solving exercises, and even well-designed paper-and-pencil scenarios can provide the essential cognitive engagement at a fraction of the cost. The military's own experience supports this — the most effective training often occurs in relatively low-fidelity environments with high-quality facilitation and debriefing.

3.5 SIMULATION, WARGAMING, AND THE DEVELOPMENT OF JUDGMENT

Military training makes extensive use of simulation at every level of fidelity — from full-motion flight simulators that cost tens of millions of dollars to tabletop wargames using paper maps and cardboard counters. Each serves a different purpose, and the relationship between them illuminates the principles of effective simulation design.

High-fidelity simulation is used primarily for developing procedural skills and for practicing emergency procedures. A full-motion flight simulator replicates the cockpit environment with sufficient accuracy that pilots can practice engine failures, instrument malfunctions, and extreme weather without risking actual aircraft. The learning is visceral — the simulator responds in real time to control inputs, and the consequences of poor decisions are immediately apparent. Pilots report that the physiological stress response to a simulated engine fire is remarkably similar to the real thing.

Medium-fidelity simulation sacrifices some physical realism for flexibility and lower cost. Part-task trainers focus on specific skills (instrument navigation, radio communication) without replicating the entire cockpit environment. These are used for focused practice on particular weaknesses identified during higher-fidelity evaluations.

Low-fidelity simulation — tabletop wargames, sand table exercises, command post exercises — uses abstraction to focus on decision-making and judgment rather than procedural skill. A wargame does not require the participants to fly the aircraft; it requires them to decide when and where to commit forces, how to respond to unexpected developments, and how to manage trade-offs between competing objectives under time pressure and uncertainty. The physical environment is deliberately simple precisely because the cognitive demands are deliberately complex.

The interesting finding from military training research is that the relationship between fidelity and learning is not linear. Increasing physical fidelity produces diminishing returns for many learning objectives, while the quality of the scenario design, the facilitation, and the debriefing are far more consequential. A well-designed tabletop exercise with skilled facilitation can produce more judgment development than a high-fidelity simulation with poor scenarios and no debriefing. This insight — that the learning is in the design and the debriefing, not in the technology — is one of the most important practical lessons from the training world.

The military's wargaming tradition is particularly relevant to judgment development because wargames inherently involve ambiguity, incomplete information, adversarial dynamics, and multiple decision points with cascading consequences. These are exactly the conditions under which judgment develops, according to the Kahneman-Klein framework. A tactical wargame does not have a single correct answer — it has better and worse decisions, and the quality of a decision depends on the reasoning behind it and the information available at the time it was made. This is far closer to real-world judgment than the well-defined problems that characterize most educational assessment.

The wargaming tradition also embeds what might be called "red team" thinking — the systematic attempt to identify how plans can fail. The military routinely assigns teams to play the adversary's role, specifically tasked with finding and exploiting weaknesses in the friendly plan. This practice develops the capacity to think about one's own assumptions critically — a metacognitive skill that maps directly onto layer 4 of the competence stack.

3.6 THE MILITARY TRAINING RESEARCH INFRASTRUCTURE

One of the least-appreciated assets of the training tradition is its research infrastructure. The U.S. Army Research Institute for the Behavioral and Social Sciences (ARI), the Navy's Office of Naval Research (ONR), the Air Force Research Laboratory (AFRL), and RAND Corporation have produced enormous amounts of research on learning, training, human factors, and organizational performance.

This research is published in venues that education researchers rarely encounter — *Military Psychology*, *Human Factors*, the proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), technical reports from ARI and RAND. The research is often of high quality — the military funds rigorous evaluations because it needs to know what works — but it exists in a separate bibliometric universe from education research.

Salas, Tannenbaum, Kraiger, and Smith-Jentsch (2012) — in what is arguably the most comprehensive review of organizational training science ever published — synthesized decades of this research into a set of evidence-based recommendations. Their key findings, published in *Psychological Science in the Public Interest* (FWCI 27.16), include:

1. **Training works.** The meta-analytic evidence consistently shows that well-designed training programs improve knowledge, skills, and job performance.
2. **Design matters.** Not all training is equally effective. The way training is designed, delivered, and implemented determines its impact.
3. **Before training matters.** Needs assessment, trainee readiness, and organizational support must be addressed before training begins.
4. **During training matters.** Active practice, feedback, scaffolding, and realistic scenarios produce better outcomes than passive instruction.
5. **After training matters.** Transfer support, follow-up, and organizational reinforcement determine whether training gains persist on the job.

These findings may seem obvious, but they represent a level of empirical discipline that much of educational practice lacks. The training world has spent decades rigorously studying what makes instruction effective — and its conclusions converge remarkably with what cognitive science and instructional design research have found independently (as documented in L1-004).

4 MEDICAL AND PROFESSIONAL TRAINING: SIMULATION, RESIDENCY, AND COMPETENCY

4.1 SIMULATION-BASED MEDICAL EDUCATION

Medicine has developed the most sophisticated training culture outside the military, and simulation is its centerpiece. McGaghie, Issenberg, Cohen, Barsuk, and Wayne (2011) conducted a meta-analytic comparative review of simulation-based medical education (SBME) with deliberate practice versus traditional clinical education. Their finding was unequivocal: SBME with deliberate practice produced superior clinical skill outcomes compared to traditional clinical education alone. The FWCI of 37.75 for this paper reflects its transformative impact on medical education.

Issenberg, McGaghie, Petrusa, Lee Gordon, and Scalese (2005) identified the features of high-fidelity medical simulations that lead to effective learning. Their review of 109 studies identified ten key features:

1. Providing feedback
2. Repetitive practice
3. Curriculum integration
4. Range of difficulty
5. Multiple learning strategies
6. Capture of clinical variation
7. Controlled environment
8. Individualized learning
9. Defined outcomes
10. Simulator validity

What is striking about this list is how completely it aligns with the principles identified by cognitive science and instructional design research. Feedback (L1-003), repetitive practice with spacing (L1-004), progressive difficulty (expertise-adaptive instruction), defined outcomes (criterion-referenced assessment) — the simulation-based medical education literature independently converged on the same design principles that the learning sciences have identified. This convergence across independent research traditions strengthens confidence in the underlying principles.

4.2 THE RESIDENCY MODEL: APPRENTICESHIP AT SCALE

Medical residency is the largest structured apprenticeship system in the developed world. Approximately 140,000 physicians in the United States are in residency training at any given time, learning to practice medicine through a model that combines graduated responsibility, supervised practice, feedback, and progressive independence.

The residency model embodies several design principles that distinguish it from classroom education:

Learning through authentic practice. Residents learn to practice medicine by practicing medicine — under supervision, with graduated responsibility, but in real clinical environments with real patients. This is not an educational philosophy; it is a practical necessity. Medical competence cannot be developed through lectures and textbooks alone because the critical knowledge is tacit — it exists in the recognition of patterns, the calibration of judgment, and the management of uncertainty that can only be developed through direct experience.

Graduated responsibility. First-year residents (interns) operate under close supervision, managing straightforward cases with attending physicians readily available. As they progress, they take on more complex cases with less supervision. By their final year, senior residents function semi-independently, with attending oversight available but not constantly present. This graduated release of responsibility mirrors the expertise-adaptive model that L1-004 identified — scaffolding is progressively faded as competence develops.

The teaching hospital culture. Teaching hospitals maintain a culture in which teaching and learning are integrated into clinical practice. Senior residents teach junior residents, attending physicians teach residents, and the expectation that everyone is both learning and teaching is woven into the organizational fabric. This creates what Lave and Wenger (1991) would recognize as a community of practice — an environment where newcomers develop competence through legitimate peripheral participation, gradually taking on more central roles as their expertise grows.

Morbidity and mortality conferences. M&M conferences — regular meetings in which adverse outcomes are reviewed by the entire department — function as medical AARs. Cases where things went wrong are analyzed in detail: what happened, what was the reasoning, what could have been done differently. Like military AARs, effective M&M conferences focus on systems and processes rather than blame, though the culture around M&M varies considerably across institutions.

4.3 COMPETENCY-BASED MEDICAL EDUCATION

In recent decades, medical education has undergone a transformation toward competency-based medical education (CBME) — a framework that defines graduation readiness in terms of demonstrated competencies rather than time served. Under CBME, a resident does not complete training after a fixed number of years; they complete training when they have demonstrated competence across a defined set of skills and judgments.

This is the medical profession's version of Bloom's mastery learning, and it represents a direct challenge to the time-based structure of traditional education. The implementation challenges are significant — defining competencies with sufficient specificity, developing valid and reliable assessment methods, managing the variable-time implications for scheduling and resources — but the conceptual alignment with the training tradition's emphasis on criterion-referenced outcomes is clear.

The milestone framework developed by the Accreditation Council for Graduate Medical Education (ACGME) provides a structured progression from novice to expert across multiple competency domains, including patient care, medical knowledge, practice-based learning, interpersonal skills, professionalism, and systems-based practice. This framework addresses all five layers of the competence stack: knowledge (medical knowledge), skill (patient care), judgment (practice-based learning, clinical reasoning), metacognition (self-assessment, identification of learning needs), and character (professionalism, ethical practice).

The CBME movement offers important lessons for education, but also important cautions. The lessons are that competency-based frameworks *can* be developed even for complex, judgment-intensive domains — medicine is not a simple domain with easily measurable outcomes. The cautions are that implementation is difficult and resource-intensive. Defining competencies with sufficient specificity to guide assessment requires enormous effort. Developing assessment methods that are both valid and reliable for judgment-level competencies (layers 3–5) remains a challenge — most CBME assessment relies on workplace-based assessment by trained observers, which is expensive and susceptible to rater variability. Managing the variable-time implications of true competency-based progression creates logistical headaches for institutions designed around fixed timelines. And there is a persistent tension between the CBME ideal (graduate when you're competent) and the institutional reality (hospitals need a predictable supply of new physicians each year).

Despite these challenges, CBME represents the most ambitious attempt to date to import the training world's mastery learning philosophy into a large-scale educational system. Its successes and failures provide a valuable evidence base for Applied Pedagogy's own implementation efforts. The most important finding from the CBME experience may be that the *direction* is clearly right — moving from time-based to competency-based assessment improves both learning and professional preparation — even when the *implementation* is messy and incomplete.

4.4 OTHER PROFESSIONAL TRAINING MODELS

Beyond medicine, several other professional training models offer relevant evidence:

Law. Legal education combines classroom instruction (law school) with supervised practice (articling, clerkships). The Socratic method used in law schools — in which professors challenge students' reasoning through structured questioning — is, in effect, a form of productive failure applied to legal reasoning. Students are forced to articulate and defend positions, have their reasoning challenged, and develop the capacity to think under pressure.

Skilled trades. Apprenticeship in skilled trades (electricians, plumbers, carpenters) follows a time-honored model of graduated responsibility under supervision, with competence assessed through practical demonstrations rather than written exams. The apprenticeship model predates formal education by centuries and embodies many of the design principles identified by the training literature: learning through authentic practice, graduated responsibility, criterion-referenced assessment, and close mentoring relationships.

Aviation beyond the cockpit. Air traffic control training is one of the most demanding training programs in the civilian world — the Federal Aviation Administration's training pipeline has historically washed out 50% or more of candidates. The program combines classroom instruction, simulation, and supervised on-the-job training with progressive complexity. Controllers train on increasingly busy and complex airspace sectors until they can manage the full complexity of their assigned position.

Engineering. Engineering education follows a path from theoretical instruction through structured design projects to internships and early career practice. The capstone design project — a team-based engineering design challenge that spans an academic term — is one of the few educational practices that approximates training-world conditions: students work on authentic problems, face real constraints, must coordinate with team members, and produce deliverables that are evaluated against functional criteria rather than abstract academic standards. The relative effectiveness of capstone experiences compared to traditional coursework provides indirect support for the training-world approach.

Accounting and professional certification. The accounting profession relies on a combination of university education and structured articling periods, culminating in professional examinations that function as criterion-referenced assessments. The CPA examination tests not just knowledge but the capacity to apply accounting principles to novel scenarios — a form of judgment assessment. The articling period provides the supervised practice component that university education alone cannot deliver.

4.5 THE COMMON THREAD ACROSS PROFESSIONAL TRAINING

What unites these diverse professional training models is a set of design principles that recur across domains despite the lack of cross-pollination between them:

1. **The endpoint is performance, not knowledge.** Every professional training system ultimately evaluates whether the trainee can *do the job*, not whether they can pass an exam about the job. This performance orientation shapes every upstream design decision.

2. **Supervised practice is the primary learning mechanism.** Across professions, the bulk of competence development occurs through doing the work under progressively decreasing supervision — not through classroom instruction. Classroom instruction builds the knowledge foundation, but it is practice that converts knowledge into capability.

3. **Assessment is pass/fail against a standard.** Professional licensing and certification exams are criterion-referenced: you meet the standard or you don't. There is no curve, no grade distribution, no ranking of candidates against each other. This creates a fundamentally different dynamic than competitive academic grading.

4. **The profession self-regulates through standards.** Each profession maintains its own competency standards, evaluation methods, and quality control mechanisms. The profession — not the university — has the final say on whether a practitioner is competent. This self-regulation is driven by the same incentive structure that drives training effectiveness: the profession bears the cost of incompetent practitioners.

5. **Mentoring relationships are built into the structure.** Apprenticeship, residency, articling, and supervised practice all embed mentoring as a structural feature, not an optional supplement. The mentor provides the feedback, modeling, and graduated challenge that formal instruction alone cannot deliver.

These principles converge with the findings from the military and aviation training literature, despite having developed independently. The convergence strengthens confidence that these are genuine principles of competence development, not artifacts of particular institutional histories.

5 WORKPLACE LEARNING: HOW COMPETENCE DEVELOPS THROUGH PRACTICE

5.1 THE INFORMAL LEARNING LITERATURE

The workplace learning literature reveals a finding that should trouble anyone who designs formal educational programs: most learning that produces workplace competence is informal. Eraut (2004) conducted a large-scale study of how early-career professionals (engineers, nurses, accountants) actually learned to do their jobs and found that the vast majority of their learning occurred through informal processes — observing colleagues, solving problems on the job, receiving feedback from supervisors and peers, and reflecting on experience. Formal training courses, while valued, accounted for a relatively small proportion of the learning that mattered for job performance.

Marsick and Watkins (2001) made a similar argument, distinguishing between *formal learning* (structured, planned, classroom-based), *informal learning* (experiential, unstructured, but intentional), and *incidental learning* (a by-product of other activities, often unconscious). They argued that informal and incidental learning are the predominant modes through which adults develop workplace competence, and that organizations that support only formal learning miss most of what matters.

5.2 THE 70-20-10 MODEL: FOLK WISDOM OR EVIDENCE?

The claim that 70% of workplace learning occurs through on-the-job experience, 20% through social interactions (coaching, mentoring, peer learning), and 10% through formal training has become almost a mantra in the corporate learning and development world. The model is typically attributed to Morgan McCall, Michael Lombardo, and Robert Eichinger at the Center for Creative Leadership in the 1980s, based on surveys of successful executives about how they learned to lead.

The evidentiary status of the 70-20-10 model is weak. The original study was based on retrospective self-reports from a small, unrepresentative sample. The specific percentages have never been rigorously validated, and there is good reason to doubt that a single ratio applies across all contexts, roles, and career stages. A surgeon's learning profile is likely very different from a software developer's, which is likely very different from a sales representative's.

However, the directional claim — that formal training is a minority component of how adults develop workplace competence, and that experiential and social learning are the majority — is well-supported by multiple lines of evidence, including Eraut's research. The practical implication is not that formal training is unimportant (it is essential for building foundational knowledge and skill) but that the design of the *work environment itself* — the affordances it provides for learning, the quality of supervision and feedback, the opportunities for graduated challenge — is a more powerful lever for competence development than any training program.

5.3 BILLETT'S WORKPLACE CURRICULUM

Stephen Billett has made the most sustained argument for understanding workplaces as learning environments in their own right — not just sites where learning from formal education is applied.

In a series of influential papers (Billett, 2001, 2002, 2006), he developed the concept of "workplace affordances" — the opportunities that workplaces provide (or fail to provide) for learning through participation.

Billett argued that workplaces have an implicit *curriculum* — a structured pathway of activities that workers engage in as they develop competence. In a well-designed workplace, newcomers begin with simpler tasks and progressively take on more complex ones as their competence grows. They receive guidance from more experienced colleagues, observe expert practice, and gradually move from peripheral to central participation. This is precisely the trajectory that Lave and Wenger (1991) described as "legitimate peripheral participation" — learning through growing involvement in a community of practice.

The quality of workplace learning, Billett argued, depends on two factors: the *affordances* the workplace provides (access to challenging tasks, guidance from experts, opportunities for practice) and the *individual's engagement* with those affordances (their willingness to participate, take initiative, seek feedback). Neither factor alone is sufficient — a workplace rich in learning opportunities benefits only workers who actively engage with them, and a highly motivated worker benefits little from a workplace that provides no opportunities for growth.

This framework has direct implications for educational design. If competence develops primarily through participation in authentic practice — through doing real work with real consequences under the guidance of more experienced practitioners — then educational programs that rely primarily on classroom instruction and decontextualized assessment are missing the primary mechanism through which competence actually develops. The implications are not that classrooms are useless but that they are insufficient — they build the foundation (layers 1–2 of the competence stack) but cannot, by themselves, develop the judgment, metacognition, and character that constitute full-stack competence.

5.4 COMMUNITIES OF PRACTICE

Lave and Wenger's (1991) *Situated Learning: Legitimate Peripheral Participation* — already foundational to the Lo survey — takes on particular significance in the training context. Their core argument is that learning is not primarily a cognitive process occurring inside individual heads but a social process of increasing participation in communities of practice.

In training contexts, communities of practice are highly visible. A fighter squadron, a surgical team, a construction crew, an engineering firm — each constitutes a community with shared practices, shared knowledge, shared standards of performance, and shared processes for integrating newcomers. The newcomer's trajectory from peripheral to central participation is not just an empirical description of how learning happens; it is the *design* of the training process. Apprenticeship, residency, mentorship, graduated responsibility — all are mechanisms for managing the newcomer's trajectory through the community of practice.

The challenge for education is that schools are communities of practice for *being a student* — not for any of the practices that students are ostensibly preparing for. The practices of schooling (sitting in lectures, taking notes, writing exams) bear little resemblance to the practices of any professional domain. This disconnect is what Ivan Illich was pointing at in his critique of schooling, and it is what motivates the training tradition's insistence on authentic practice.

5.5 ERAUT'S TYPOLOGY OF WORKPLACE KNOWLEDGE

Eraut's research contributed a particularly useful typology of the knowledge that professionals actually use in practice. He distinguished between:

Codified knowledge — formal, explicit knowledge that can be written down and taught in classrooms. Textbook knowledge, theories, procedures, regulations. This is what formal education primarily transmits.

Personal knowledge — knowledge that individuals construct through experience. This includes interpretations of situations, understanding of how things work in practice (which often differs from how they work in theory), and tacit knowledge that the practitioner may not be able to articulate.

Cultural knowledge — shared knowledge embedded in the practices and norms of a community. How things are done "around here," what counts as acceptable performance, what shortcuts are tolerated and which are not. This knowledge is transmitted through participation in the community, not through instruction.

Eraut found that early-career professionals relied primarily on codified knowledge when they first entered the workplace, but that this knowledge was rapidly supplemented and often overridden by personal and cultural knowledge as they gained experience. The implication is profound: formal education equips learners with codified knowledge, but this is only one of three types of knowledge needed for competent practice, and arguably the least important one for experienced practitioners.

This maps directly onto the competence stack. Codified knowledge corresponds to layer 1 (domain knowledge). Personal knowledge corresponds to layers 2 and 3 (skill and judgment). Cultural knowledge corresponds to layer 5 (character and disposition — the epistemic habits that are normative within a community). Formal education addresses layer 1; workplace learning addresses layers 2, 3, and 5. Layer 4 (metacognition) develops at the intersection, through structured reflection on the gap between codified knowledge and practical experience.

5.6 THE IMPLICATIONS OF WORKPLACE LEARNING FOR EDUCATIONAL DESIGN

The workplace learning literature delivers an uncomfortable message to educators: most of the competence that matters for professional practice is not developed through formal instruction. Formal education provides the necessary foundation — the codified knowledge base, the basic skills, the initial mental models — but the transformation from knowledgeable person to competent practitioner occurs primarily through practice, social interaction, and experience.

This does not mean that formal education is unimportant. A surgeon who has not learned anatomy cannot develop surgical skill through practice alone. An engineer who has not learned physics cannot develop engineering judgment through experience alone. The foundation matters. But the foundation is necessary and insufficient — like learning the rules of chess without ever playing a game.

The practical implication for Applied Pedagogy is that curriculum design must attend to the *full ecosystem* of learning, not just the formal instruction component. This means designing not only lessons and assessments but also practice environments, mentoring structures, feedback systems, and communities of practice that support the development of personal and cultural knowledge alongside codified knowledge.

Part III

TRANSFER, RELIABILITY, AND COMPETENCE

6 TRANSFER OF TRAINING: WHEN AND WHY IT WORKS (OR DOESN'T)

The transfer problem — whether what is learned in one context can be applied in another — is central to both training and education. But the training world approaches it differently: as a practical design problem with real financial consequences, not primarily as a theoretical puzzle.

6.1 THE BALDWIN AND FORD FRAMEWORK

Baldwin and Ford's (1988) foundational review of transfer of training established a framework that has organized the field for nearly four decades. They identified three categories of factors that influence transfer:

1. **Training inputs:** trainee characteristics (ability, personality, motivation), training design (learning principles, sequencing, content relevance), and work environment characteristics (support, opportunity to use).
2. **Training outputs:** learning and retention during training.
3. **Conditions of transfer:** generalization of learned material to the job context and maintenance of trained skills over time.

Their key insight was that transfer depends not only on the quality of the training itself but also on what happens *before* training (trainee preparation and motivation) and *after* training (organizational support for application). Training that produces excellent learning in the classroom but is not supported by the work environment will not transfer. This systemic perspective — that transfer is an organizational problem, not just an instructional one — distinguishes the training tradition from education, where transfer is typically framed as a property of the learner or the instruction.

6.2 THE BLUME META-ANALYSIS

Blume, Ford, Baldwin, and Huang (2010) conducted the definitive meta-analysis of transfer of training, examining 89 empirical studies. Their findings confirmed and extended Baldwin and Ford's framework:

- **Cognitive ability** was the strongest individual predictor of transfer ($\rho = .37$), confirming that more capable learners transfer more.
- **Motivation to transfer** was a significant predictor, suggesting that the learner's intention to apply what they have learned matters.
- **Supervisor support** and **opportunity to perform** were the strongest environmental predictors — confirming that the organizational context is at least as important as the training itself.

- **Training design** variables (particularly the use of realistic practice and error management training) were significant predictors.
- **Open skills** (complex, principle-based) transferred less readily than **closed skills** (procedural, algorithmic), a finding that echoes the well-structured vs. ill-structured domain distinction from L1-004.

Burke and Hutchins (2007) conducted a complementary integrative review and reached similar conclusions, emphasizing that transfer is a multi-level phenomenon requiring intervention at the individual, instructional, and organizational levels.

6.3 ERROR MANAGEMENT TRAINING

One of the most innovative findings from the transfer literature is the effectiveness of error management training (EMT). Keith and Frese (2008) found that training programs that *encourage* learners to make errors during practice — and then help them learn from those errors — produce superior transfer compared to error-avoidant training that tries to prevent mistakes.

This finding is directly parallel to Kapur's (2024) productive failure research in education, which L1-004 explored in depth. The mechanisms are similar: errors during practice activate deeper processing, develop diagnostic skills, and build a richer understanding of the problem space. Error management training also develops metacognitive skills — learners who have experienced and recovered from errors are better at monitoring their own performance and detecting problems early.

The error management training literature adds an important nuance to the productive failure framework: the organizational context matters. Error management training works best in organizations with an error management culture — one that treats errors as learning opportunities rather than as evidence of incompetence (van Dyck, Frese, Baer, & Sonnentag, 2005). In organizations with a blame culture, encouraging errors during training may be counterproductive because learners will be anxious rather than exploratory.

6.4 THE RELAPSE PREVENTION MODEL

One of the more innovative approaches to sustaining training transfer comes from an unexpected source: addiction psychology. Marx (1982) adapted Marlatt and Gordon's relapse prevention model — originally developed for substance abuse treatment — to the problem of training transfer. The basic insight is powerful: trainees, like recovering addicts, face a high-risk environment when they return to the workplace. Old habits reassert themselves. The social environment may not support the new behaviors. Without specific coping strategies, "relapse" to pre-training behavior is likely.

The relapse prevention model applied to training involves three steps:

1. **Anticipating high-risk situations.** Before leaving training, identify the specific situations in which the new skills are most likely to break down — time pressure, unsupportive supervisors, ambiguous situations where the old approach feels more comfortable.
2. **Developing coping strategies.** For each high-risk situation, develop a specific plan for how to apply the new skills. "When my supervisor pressures me to take shortcuts, I will…"
3. **Managing setbacks.** Recognize that occasional failure to apply new skills is normal, not a sign that the training failed. Develop strategies for getting back on track after a setback rather than giving up entirely.

This approach to transfer is notable because it treats transfer as a *self-regulation* problem, not just an instructional design problem. It connects directly to L1-002's finding that self-regulation can be taught explicitly, and it suggests that training programs should include specific transfer planning as a standard component.

6.5 THE ORGANIZATIONAL SUPPORT FACTOR

Perhaps the most sobering finding from the transfer literature is that the most powerful predictors of whether training transfers are not properties of the training itself but properties of the organizational environment. Blume et al.'s (2010) meta-analysis found that supervisor support and opportunity to perform were among the strongest predictors of transfer — stronger than most training design variables.

This means that an excellently designed training program delivered into an unsupportive organizational environment will produce limited transfer. Conversely, even a mediocre training program may produce reasonable transfer if the organization actively supports application. The implication is that transfer is not solely an instructional design problem — it is an organizational design problem.

For training practitioners, this finding has led to the concept of "transfer climate" — the overall organizational support for applying trained skills. Transfer climate includes supervisor encouragement, peer support, opportunity to practice, absence of punishment for errors, and reward for using new skills. Organizations that invest in training without investing in transfer climate are wasting their training investment.

For educational contexts, the parallel is striking. Educational programs that prepare students well but then release them into environments that do not support the application of their learning should expect poor outcomes. The "transfer" of educational learning to professional practice depends not just on the quality of the education but on the quality of the practice environment — exactly the insight that the workplace learning literature delivers from a different direction.

6.6 WHAT THE TRANSFER LITERATURE MEANS FOR EDUCATION

The training transfer literature has several implications that education has been slow to absorb:

Context matters as much as content. The organizational environment in which learning is applied — supervisor support, opportunity to practice, cultural norms around learning — is at least as important as the quality of the instruction itself. Educational programs that prepare students excellently but then release them into unsupportive environments should expect poor transfer.

Near transfer is reliable; far transfer is rare. The transfer literature consistently shows that people transfer learning most readily to situations similar to the training context. Transfer to dissimilar contexts (far transfer) is much more difficult and less reliable. This finding converges with the conclusion from L0-001v2's Gap 1: far transfer is "the holy grail of education research" — universally desired, poorly understood, and perhaps inherently limited.

Design for transfer from the beginning. The most effective training programs design for transfer from the start — using realistic scenarios, varied practice contexts, and explicit instruction on when and how to apply learned principles. Transfer is not something that happens automatically after learning; it must be designed into the instructional system.

7 HIGH-RELIABILITY ORGANIZATIONS AND TRAINING CULTURE

7.1 WHAT MAKES HROS DIFFERENT

High-reliability organizations (HROs) — nuclear power plants, aircraft carriers, air traffic control systems, emergency departments — are organizations that operate complex, hazardous technologies with remarkably low failure rates. Karl Weick and Kathleen Sutcliffe (2007) identified five principles that characterize HRO culture in their influential *Managing the Unexpected*:

1. **Preoccupation with failure.** HROs treat every small error and near-miss as a signal that something is wrong with the system, not as a routine fluctuation to be ignored. They actively look for problems rather than waiting for problems to find them.
2. **Reluctance to simplify.** HROs resist the temptation to reduce complex situations to simple categories. They encourage nuanced understanding and are suspicious of oversimplified explanations.
3. **Sensitivity to operations.** HROs maintain continuous awareness of what is actually happening on the front lines, rather than relying on abstract reports and metrics.
4. **Commitment to resilience.** HROs invest in the capacity to detect and recover from errors quickly, recognizing that errors will inevitably occur.
5. **Deference to expertise.** When problems arise, HROs defer to the person with the most relevant expertise, regardless of their position in the hierarchy.

These principles describe not just an organizational culture but a *training culture* — a set of values and practices around how the organization develops and maintains competence. Every one of these principles has direct implications for how learning is designed and supported.

7.2 ERROR MANAGEMENT CULTURE

Van Dyck, Frese, Baer, and Sonnentag (2005) studied error management culture — the shared beliefs, norms, and practices around how an organization handles errors. They found that organizations with a stronger error management culture (treating errors as learning opportunities, communicating about errors openly, helping each other after errors) had better organizational performance than those with error aversion cultures (trying to prevent errors, punishing mistakes, covering up failures).

This finding connects directly to COMPETENCE-TARGET.md's environmental multiplier. An environment that punishes error does not just suppress error reporting — it degrades the capacity to perceive and learn from errors over time. An error management culture, by contrast, creates the feedback loops necessary for continuous improvement at all levels of the competence stack.

7.3 PSYCHOLOGICAL SAFETY AS TRAINING INFRASTRUCTURE

Edmondson's (1999) research on psychological safety, already central to L1-009's analysis, takes on additional significance in the training context. Edmondson found that teams with higher psychological safety engaged in more learning behavior — and that this was true even after controlling for team effectiveness. Psychological safety was not just a nice-to-have; it was a prerequisite for team learning.

In training contexts, psychological safety is literally part of the training infrastructure. CRM training explicitly builds psychological safety into flight crews. After-action reviews create psychologically safe spaces through rank neutrality and focus on learning rather than blame. Simulation-based training provides a safe environment for making mistakes — the simulator does not crash real aircraft.

The educational implication is that psychological safety is not merely an aspect of classroom climate — it is a first-order determinant of whether upper-layer competence can develop. An educational environment that punishes errors, rewards confident performance over honest self-assessment, or creates anxiety around evaluation is not just unpleasant — it is structurally incapable of producing the judgment, metacognition, and intellectual honesty that constitute full-stack competence.

7.4 REASON'S SWISS CHEESE MODEL AND ERROR MANAGEMENT

James Reason's (1990) *Human Error* introduced the Swiss cheese model of accident causation, which has become the foundational framework for understanding how errors propagate in complex systems. Reason argued that complex systems have multiple defensive layers (like slices of Swiss cheese), each with holes (latent conditions and active failures). An accident occurs when the holes in multiple layers align, allowing a hazard to pass through all defenses.

This model transformed how HROs think about training. Rather than focusing on preventing individual errors — which Reason argued is futile because humans are inherently fallible — HROs focus on building robust defenses and ensuring that individual errors do not propagate into system failures. Training in HROs therefore addresses multiple levels:

Individual skill training ensures that people can perform their tasks correctly under normal conditions. This is the baseline — necessary but insufficient.

Error detection training ensures that people can recognize when something has gone wrong, in themselves or in others. Mutual monitoring — the practice of watching each other's work and speaking up when you notice a problem — is a trained skill, not a natural behavior. CRM training explicitly develops this capacity.

Error recovery training ensures that when errors are detected, people know how to correct them before they propagate. This includes specific procedures (checklists, escalation protocols) and general skills (communication, coordination, assertiveness).

System design for error tolerance ensures that the system itself can absorb individual errors without catastrophic failure. This is a design principle, not a training objective — but understanding how system design affects error propagation is itself a training objective for operators and managers.

The implications for education are significant. Schools typically focus exclusively on individual skill training (layer 1 of Reason's framework) — teaching students to perform correctly. They rarely train error detection (recognizing one's own mistakes), error recovery (knowing what to do when something goes wrong), or system-level thinking (understanding how individual actions affect the larger system). Training contexts address all four levels because they cannot afford to ignore any of them.

7.5 THE CHRONIC UNEASE PRINCIPLE

One of the more counterintuitive principles from HRO research is the concept of "chronic unease" — the deliberate maintenance of a low level of anxiety about the possibility of failure, even when things are going well. Weick and Sutcliffe (2007) described this as "preoccupation with failure" — the organizational habit of treating success not as evidence that the system is safe but as evidence that the organization has not yet discovered its vulnerabilities.

This principle runs counter to the educational emphasis on positive reinforcement, growth mindset, and celebrating success. In HROs, excessive confidence is treated as a threat — because confidence reduces vigilance, and reduced vigilance allows latent failures to accumulate undetected. The ideal mental state for an HRO operator is not confident mastery but *alert competence* — the capacity to perform effectively while remaining attentive to the possibility that something might go wrong.

For the competence stack, chronic unease maps directly onto layer 4 (metacognition) and layer 5 (character). It requires the metacognitive capacity to monitor one's own confidence and recognize when confidence is outpacing evidence. It requires the character to resist the social pressure to project certainty and instead maintain a stance of humble vigilance. These are dispositions that develop through repeated experience in environments that reward them — exactly the environmental design principle that L1-009 identified.

7.6 THE BAKER-DAY-SALAS FRAMEWORK FOR TEAMWORK IN HROS

Baker, Day, and Salas (2006) proposed a framework for understanding teamwork as an essential component of high-reliability organizations. They argued that HRO performance depends not just on individual competence but on team competence — the capacity of the team as a system to coordinate, communicate, adapt, and recover from errors.

Their framework identifies five core components of effective teamwork: team leadership, mutual performance monitoring, backup behavior, adaptability, and team orientation. Critically, these are *trainable* competencies — CRM training, TeamSTEPPS (Team Strategies and Tools to Enhance Performance and Patient Safety), and other team training programs have demonstrated that teams can be trained to work more effectively together (Salas, DiazGranados, Klein, Burke, Stagl, Goodwin, & Halpin, 2008). The meta-analysis of team training effectiveness found a medium-to-large effect ($d = 0.34$ for performance outcomes), with effects moderated by team stability and training content.

8 The Competence Stack in Training Contexts

One of the central questions in this investigation is how training contexts address the five layers of the competence stack — and particularly whether they do a better job with layers 3–5 (judgment, metacognition, character) than traditional education.

8.1 Layer 1: Domain Knowledge

Training contexts build domain knowledge, but they do so differently from education. Rather than front-loading knowledge through lectures and textbooks, training typically provides knowledge on a just-in-time basis — delivering the information learners need when they need it for the task at hand. This approach aligns with van Merriënboer's 4C/ID model (as L1-004 documented), which distinguishes between "supportive information" (provided before and during learning tasks) and "procedural information" (provided just-in-time during practice).

The military's emphasis on task analysis means that knowledge is always organized around performance requirements rather than disciplinary structures. A pilot does not study aerodynamics as a subject and then apply it to flying; they study the aerodynamic principles needed to fly the aircraft as part of learning to fly the aircraft. This performance orientation ensures that knowledge is always connected to its application — addressing the inert knowledge problem that plagues much of formal education.

8.2 Layer 2: Skill

Training contexts excel at skill development because they provide extensive deliberate practice with feedback. The simulation literature demonstrates that deliberate practice in simulated environments produces skill acquisition that transfers to real performance (McGaghie et al., 2011). The key features are clear performance standards, immediate feedback, repetitive practice at an appropriate level of challenge, and progressive complexity.

The training world has also identified the critical role of *automaticity* — the development of procedural fluency that frees cognitive resources for higher-level tasks. Pilots practice emergency procedures until they become automatic, so that in an actual emergency, cognitive resources are available for the non-routine decision-making that the specific situation requires. This understanding of automaticity as a foundation for judgment (rather than a substitute for it) is important for educational design. The current backlash against "rote learning" and "drill" in education sometimes fails to recognize that procedural fluency is a necessary (though not sufficient) foundation for higher-level competence.

8.3 Layer 3: Judgment

This is where training contexts may have the strongest advantage over education. Judgment — the capacity to determine which knowledge and skills to deploy in a given situation — develops through exposure to varied, consequential, and ambiguous situations with feedback. Training contexts routinely provide exactly this:

- **Flight simulation** presents pilots with novel emergency scenarios that require rapid judgment under uncertainty. No two scenarios are identical, and the pilot must decide which procedures to follow, when to deviate from standard procedures, and how to prioritize competing demands.
- **Military wargaming** presents officers with complex tactical and strategic situations that do not have clear right answers. Participants must make decisions with incomplete information, manage trade-offs between competing objectives, and deal with the consequences of their decisions.
- **Medical simulation** presents clinicians with complex patient presentations that require differential diagnosis — the systematic consideration and elimination of possible diagnoses. This is judgment training in its purest form.
- **After-action reviews** develop judgment retrospectively by systematically comparing expected outcomes to actual outcomes and analyzing why they differed.

The Kahneman-Klein framework for developing intuitive expertise (Kahneman & Klein, 2009) — explored in L1-009 — specifies that valid intuitive judgment can develop in domains that are both high-validity (regularities exist to be learned) and high-feedback (outcomes are visible and timely). Training contexts in aviation, medicine, and the military typically satisfy both conditions. Most educational contexts satisfy neither.

8.4 LAYER 4: METACOGNITION

Training contexts develop metacognition through several mechanisms:

Calibration through prediction and feedback. Pilots are trained to anticipate aircraft behavior, form mental models of situations, and then check those models against reality. The simulation environment provides immediate feedback on the accuracy of their mental models — if they predicted the aircraft would behave one way and it behaved differently, they know their model was wrong. This is metacognitive training embedded in practice.

Structured reflection. AARs and debriefing sessions require participants to examine their own cognitive processes — what they were thinking, what information they were attending to, what assumptions they were making. This reflective practice develops the self-monitoring capacity that is the hallmark of metacognition.

Checklists as metacognitive scaffolds. The aviation industry's extensive use of checklists functions as a metacognitive scaffold — a external tool that compensates for the limitations of human memory and attention. Gawande's (2009) *The Checklist Manifesto* documented how this simple technology, transferred from aviation to surgery, significantly reduced complications and mortality. Checklists do not replace judgment; they ensure that routine steps are not overlooked, freeing cognitive resources for the judgment-intensive aspects of the task.

8.5 LAYER 5: CHARACTER AND DISPOSITION

This is the most provocative claim: that training contexts may develop epistemic character — intellectual honesty, tolerance for uncertainty, courage to deliver bad news — more effectively than educational contexts. The evidence is indirect but suggestive:

Error management culture. Organizations that treat errors as information rather than as evidence of incompetence cultivate the disposition to engage honestly with reality. Over time, participants in such cultures develop the habit of reporting errors, acknowledging uncertainty, and seeking feedback — not because they are told to, but because the environment rewards it.

Rank neutrality in AARs. The norm that junior participants can and should challenge senior participants' decisions during AARs cultivates the courage to speak truth to power. This norm must be actively maintained — it is fragile and can be destroyed by a single senior leader who punishes honest feedback — but when it works, it creates an environment in which intellectual honesty is habitual rather than heroic.

The "sterile cockpit" rule. FAA regulations prohibit non-essential conversation during critical phases of flight. This rule embodies a cultural commitment to prioritizing reality over social comfort — during critical operations, the only thing that matters is accurate information about the state of the aircraft and its environment. This is a structural intervention that reinforces the disposition to attend to reality rather than to social dynamics.

Psychological safety research. Edmondson's work demonstrates that environments with high psychological safety produce more learning behavior, more innovation, and better performance. The implication for character development is that intellectual honesty is not primarily a personality trait to be selected for but an environmental outcome to be designed for — consistent with L1-009's finding that layer 5 is primarily an environmental design problem.

The Sexton cross-sectional study. Sexton, Thomas, and Helmreich (2000) conducted a cross-sectional survey comparing attitudes toward error and teamwork in aviation and medicine — and the findings illuminate how far medicine had yet to travel on the CRM path. They found that surgeons were significantly more likely than pilots to deny the effects of stress and fatigue on performance. While 97% of cockpit crew members endorsed the importance of flat hierarchies and open communication, only 55% of surgical staff did. This difference reflects the respective maturity of CRM culture in the two domains — aviation had been working on these cultural norms for twenty years; medicine was just beginning.

The practical implication is that the training culture's advantages at layer 5 are not automatic. They are achieved through sustained, deliberate cultural intervention — through programs like CRM that explicitly address authority gradients, communication norms, and error management. Without such programs, even high-stakes training environments can fail to develop the epistemic honesty that layer 5 requires.

However, it is important to acknowledge what training contexts do *not* do well at layer 5. Military training cultures, for all their emphasis on honest self-assessment within the training context, can also cultivate conformity, deference to authority, and resistance to dissent outside the training context. The "rank neutrality" of the AAR is an exception to the military's normal hierarchical culture, not a reflection of it. Medical training cultures have historically included elements of hazing, sleep deprivation, and emotional abuse that are antithetical to the development of epistemic honesty. The training world's advantages at layer 5 are real but partial, and they are most evident in specific practices (AARs, CRM, simulation debriefing) rather than in the overall culture of training institutions.

Part IV

SYNTHESIS

9 WHAT EDUCATION COULD IMPORT: TRANSFERABLE DESIGN PRINCIPLES

The central practical question of this investigation is: what can education learn from training? Not all training principles transfer to educational contexts — the structural differences are real — but several design principles are both evidence-based and adaptable.

9.1 PRINCIPLE 1: CRITERION-REFERENCED ASSESSMENT AS THE DEFAULT

The training practice: Assessment determines whether the learner can meet the performance standard, not how they rank relative to other learners. Training continues until the standard is met.

The evidence: Mastery learning meta-analyses (Kulik et al., 1990) consistently show large effect sizes ($d = 0.5\text{–}1.0$) when time constraints are relaxed and criterion-referenced approaches are used. The CBME movement in medical education provides a large-scale implementation model.

The educational adaptation: Design assessment around clearly defined competencies. Replace norm-referenced grading with criterion-referenced assessment wherever institutional constraints permit. Allow flexible time-to-mastery where possible. When external constraints require fixed time (semester systems), use criterion-referenced assessment within those constraints — students know exactly what they need to demonstrate, and assessment provides information about progress toward that standard rather than a ranking among peers.

Connection to prior agents: This aligns with L1-003's recommendation for standards-based reporting and L1-002's finding that controlling assessment undermines motivation. Criterion-referenced assessment is inherently more autonomy-supportive because the standard is transparent and achievable.

9.2 PRINCIPLE 2: SYSTEMATIC DEBRIEFING AFTER COMPLEX LEARNING ACTIVITIES

The training practice: After every significant training event, conduct a structured debriefing. What was supposed to happen? What actually happened? Why? What will we do differently?

The evidence: Tannenbaum and Cerasoli (2012) found $d \approx 0.67$ for debriefing effectiveness, with facilitated debriefs outperforming unfacilitated ones.

The educational adaptation: Build structured reflection into the curriculum after complex learning activities — projects, simulations, lab work, case analyses. Train instructors in facilitation skills. Create norms of honest self-assessment (what went well, what didn't, why). Focus on processes and decisions, not just outcomes.

Connection to prior agents: This aligns with L1-009's recommendation for after-action reviews as a mechanism for developing judgment and metacognition.

9.3 PRINCIPLE 3: SIMULATION AND SCENARIO-BASED LEARNING FOR JUDGMENT DEVELOPMENT

The training practice: Use simulations — from high-fidelity flight simulators to low-fidelity role-playing exercises — to expose learners to varied, challenging, ambiguous situations that require judgment.

The evidence: McGaghie et al. (2011) demonstrated superiority of simulation-based training. Gaba (2004) argued that simulation effectiveness depends on cognitive fidelity, not physical fidelity.

The educational adaptation: Design scenario-based learning activities that present students with complex, ambiguous problems requiring judgment — case studies, role-playing, simulated decision-making exercises. The key is not expensive equipment but carefully designed scenarios that require the cognitive processes being developed, followed by structured debriefing.

Connection to prior agents: This aligns with L1-009's recommendation for developing judgment through varied, consequential experience and L1-004's emphasis on productive failure and guided inquiry for intermediatelearners.

9.4 PRINCIPLE 4: ERROR MANAGEMENT RATHER THAN ERROR PREVENTION

The training practice: Design practice environments that encourage errors and use them as learning opportunities. Create psychological safety around errors during training while maintaining high standards for performance.

The evidence: Keith and Frese (2008) found error management training produces superior transfer. Van Dyck et al. (2005) found error management culture improves organizational performance.

The educational adaptation: Normalize errors during practice. Create a clear distinction between practice (where errors are expected and informative) and evaluation (where performance is assessed). Build error analysis into the curriculum — when students make errors, mine them for learning value rather than simply marking them wrong.

Connection to prior agents: This aligns with L1-004's emphasis on productive failure (Kapur, 2024) and L1-003's recommendation to normalize failure and productive struggle.

9.5 PRINCIPLE 5: DESIGN THE LEARNING ENVIRONMENT BEFORE THE CURRICULUM

The training practice: HROs invest heavily in organizational culture — psychological safety, error management culture, deference to expertise — as the foundation for effective performance and learning.

The evidence: Edmondson (1999) demonstrated that psychological safety is a prerequisite for team learning. Van Dyck et al. (2005) showed error management culture predicts performance. The convergence of evidence from multiple L1 agents (L1-002, L1-003, L1-009) points in the same direction.

The educational adaptation: Before optimizing content or instructional methods, design the learning environment. Establish norms of psychological safety, honest self-assessment, and error-as-information. Model intellectual honesty from positions of authority. Create structures that reward accuracy over confidence.

Connection to prior agents: This is the central recommendation of L1-009 — environment first.

9.6 Principle 6: Feedback That Is Fast, Specific, and Actionable

The training practice: Training environments provide immediate, specific feedback connected to performance outcomes. Feedback is diagnostic (what needs to improve) rather than evaluative (how you rank).

The evidence: Salas et al. (2012) identified feedback as one of the most critical design features. L1-003 documented the evidence for task-level feedback and the harm of delayed, evaluative feedback.

The educational adaptation: Minimize feedback latency. Make feedback specific to the task and the process, not the person. Connect feedback to the next learning activity — every piece of feedback should be followed by an opportunity to act on it.

9.7 Principle 7: Graduated Responsibility with Progressive Independence

The training practice: Newcomers begin with simpler tasks under close supervision and progressively take on more complex tasks with less supervision as their competence develops. The endpoint is independent performance.

The evidence: This is the universal structure of apprenticeship, residency, and progressive training systems. It aligns with the expertise-adaptive model from L1-004 and with Vygotsky's zone of proximal development.

The educational adaptation: Design learning sequences with explicit scaffolding reduction. Begin with worked examples and high support, gradually releasing responsibility to the learner. The trajectory should be visible to the learner — they should be able to see themselves moving toward independence.

9.8 Principle 8: Training the Team, Not Just the Individual

The training practice: CRM, TeamSTEPPS, and other team training programs recognize that competence has a social dimension. Teams are trained to communicate, coordinate, and support each other as a system.

The evidence: Salas et al. (2008) found a medium-to-large effect ($d = 0.34$) for team training on performance outcomes. CRM has been associated with dramatic improvements in aviation safety.

The educational adaptation: Design collaborative learning activities that develop team competencies — communication, mutual monitoring, backup behavior — alongside content knowledge. Train students in productive disagreement, active listening, and constructive feedback. Assess group processes, not just group products.

9.9 Principle 9: Checklists as Metacognitive Scaffolds

The training practice: Aviation and surgery use checklists to ensure that routine steps are not overlooked under stress or time pressure, freeing cognitive resources for judgment-intensive decisions.

The evidence: Gawande (2009) documented dramatic reductions in surgical complications when a simple checklist was implemented. The aviation industry's safety record is built partly on extensive use of standardized checklists for every phase of flight.

The educational adaptation: Develop domain-specific checklists that scaffold the metacognitive process of monitoring one's own work. These are not "dumbed down" aids — they are cognitive tools that even experts use. A writing checklist that asks "Have I considered the strongest counterargument?" or a problem-solving checklist that asks "Have I verified my answer against the original problem?" externalizes metacognitive processes until they become internalized habits.

9.10 PRINCIPLE 10: RED TEAM THINKING FOR INTELLECTUAL RIGOR

The training practice: Military planning routinely assigns a team to play the adversary's role, specifically tasked with finding and exploiting weaknesses in the plan. This institutionalized skepticism prevents groupthink and overconfidence.

The evidence: The practice is widespread in military and intelligence communities and has been adopted by some business organizations. Formal evaluation is limited, but the practice aligns with research on cognitive biases (confirmation bias, groupthink) and the demonstrated effectiveness of structured devil's advocacy.

The educational adaptation: Build structured adversarial thinking into the curriculum. When students develop arguments, require them to develop the strongest counterargument before finalizing their position. In group work, assign a "red team" role that is explicitly tasked with finding weaknesses in the group's approach. This develops the capacity to think critically about one's own ideas — a metacognitive skill that is difficult to develop without structural support.

10 HONEST ASSESSMENT: CONFIDENCE LEVELS AND OPEN QUESTIONS

10.1 WHAT WE CAN SAY WITH HIGH CONFIDENCE

Training works, and design matters. The meta-analytic evidence is overwhelming that well-designed training programs improve knowledge, skills, and job performance (Salas et al., 2012; Aguinis & Kraiger, 2009). This is one of the most robust findings in all of organizational science. The effect is moderated by training design features — programs that incorporate active practice, feedback, realistic scenarios, and transfer support produce better outcomes than those that rely on passive instruction.

Simulation-based training is effective. The evidence that simulation-based training with deliberate practice produces superior outcomes compared to traditional instruction is strong, particularly in medical education (McGaghie et al., 2011) but also across other domains.

Debriefing enhances learning. The meta-analytic evidence for the effectiveness of structured debriefing ($d \approx 0.67$) is robust (Tannenbaum & Cerasoli, 2012).

Psychological safety is a prerequisite for learning. Edmondson's (1999) finding has been replicated across contexts and is one of the most highly-cited findings in organizational research (10,000+ citations, FWCI 42.93).

Criterion-referenced assessment produces more learning than norm-referenced assessment when time constraints are relaxed. The mastery learning evidence supports this (Bloom, 1968; Kulik et al., 1990).

Transfer depends on organizational context, not just training quality. The meta-analytic evidence (Blume et al., 2010) clearly shows that supervisor support and opportunity to perform are critical for transfer.

10.2 WHAT WE CAN SAY WITH MEDIUM CONFIDENCE

CRM training has contributed to improved aviation safety. The causal link between CRM training and the reduction in aviation accidents is plausible and consistent with the evidence, but cannot be definitively established because many other factors (improved aircraft technology, better regulations, improved air traffic control) have changed simultaneously (Salas et al., 2001, 2006).

Error management training produces superior transfer. The evidence is promising but based on a relatively small number of studies. The conditions under which error management training works (supportive organizational culture, appropriate task complexity) need further specification.

The 70-20-10 ratio is directionally correct. Most workplace learning is informal, but the specific percentages have no rigorous empirical basis. The ratio likely varies substantially across contexts.

Training contexts develop upper-layer competence more effectively than schools. This is the central claim of this investigation, and while the evidence is consistent with it, a direct, controlled comparison has never been conducted. The structural arguments are compelling, but the empirical evidence is largely indirect.

10.3 WHAT WE DON'T KNOW

Whether training design principles transfer cleanly to educational contexts. The structural differences between training and education (different populations, different purposes, different institutional constraints, different time scales) mean that principles that work in training may not work — or may work differently — in education. The evidence base for importing training principles into K-12 and higher education is thin.

How to scale training-world practices in educational contexts. AARs work partly because training events have clear objectives and measurable outcomes. How do you debrief a literature class? Criterion-referenced assessment works partly because training has clear performance standards. How do you define criterion standards for critical thinking or creative writing?

The long-term effects of training-style approaches. Training research typically measures short-term outcomes — did the trainee learn the skill? Did it transfer to the job? The long-term developmental effects of training-style approaches on motivation, identity, and lifelong learning are almost entirely unstudied.

The interaction between training culture and individual development. Military training culture produces rapid competence acquisition, but it also produces conformity, obedience, and (in some contexts) psychological harm. The training world's practices cannot be imported without careful attention to which cultural elements are necessary for effectiveness and which are incidental or harmful.

Whether the incentive structure can be artificially created. The training world's effectiveness may be inseparable from its incentive structure — the fact that the organization bears real costs for failure. Can educational institutions create analogous incentive structures without the existential stakes? Or does the effectiveness of training-world practices depend on the very conditions that make them necessary?

11

THE STRUCTURAL INSIGHT REVISITED

This investigation began with a structural observation: learning works differently when the institution doing the training bears the cost of failure. The evidence supports this observation and helps specify exactly how it works differently.

When the institution pays for failure:

- **Assessment becomes diagnostic** because the institution needs to know what the learner can actually do, not how they rank against peers.
- **Time becomes flexible** because releasing an incompetent performer is costlier than extending training.
- **Feedback becomes fast and specific** because the institution cannot afford the learning loss from slow, vague feedback.
- **Errors during practice are tolerated** because the institution understands that errors are how learning happens, and would rather errors occur in training than on the job.
- **The learning environment is designed** because the institution recognizes that culture, norms, and structures affect learning as much as content and instruction.
- **The full competence stack is addressed** because the institution needs people who can exercise judgment, monitor their own performance, and tell the truth under pressure — not just people who can recall facts and execute procedures.

These are not mysterious insights. They are the logical consequences of an incentive structure that aligns the institution's interests with the learner's development. The tragedy of formal education is that its incentive structure often works against learning — schools are rewarded for processing students efficiently (time-constant, learning-variable), for differentiating students (norm-referenced assessment), and for maintaining institutional authority (which discourages the psychological safety necessary for upper-layer competence development).

Applied Pedagogy's opportunity is to adopt training-world design principles within an educational context — to build learning environments that provide criterion-referenced assessment, fast and specific feedback, graduated responsibility, structured debriefing, error management culture, and psychological safety. The evidence says these practices work. The challenge is implementing them in a context that lacks the existential stakes that force training organizations to get it right.

There is a temptation to romanticize the training world — to present it as a utopia of evidence-based practice where everything works because the stakes are real. This would be a mistake. The training world has its own pathologies: bureaucratic rigidity, resistance to innovation, hazing and abuse, excessive emphasis on compliance over creativity, and the tendency to confuse having been through a training program with actually having developed competence. The U.S. military, for all its sophisticated training infrastructure, is also capable of producing PowerPoint culture, checkbox compliance, and performative training that develops nothing except the ability to pass a certification test.

What the training world offers is not a model to be copied wholesale but a *natural experiment* to be studied. The structural conditions of training — skin in the game, criterion-referenced assessment, flexible time, tight feedback loops, real consequences — create an environment in which the principles of effective learning have been tested and refined under conditions of genuine accountability. These principles are not unique to training; they are the same principles that cognitive science, instructional design, and motivation research have identified from other directions. The training world's contribution is not to discover new principles but to demonstrate that these principles actually work when implemented with the rigor that real stakes demand.

The PI's observation that training works better than education is supported by the evidence — but with an important caveat. Training works better at producing *competent performers in defined domains.* Whether training-world approaches can produce the broader outcomes that education claims to pursue — intellectual curiosity, democratic citizenship, aesthetic appreciation, personal flourishing — is an open question that this investigation cannot answer. What it can say is that the training world's design principles for developing competence are rigorously evidence-based, and that education ignores them at its own cost.

11.1 A FINAL NOTE ON THE CONVERGENCE

This investigation — the sixth Level 1 dissertation in the Applied Pedagogy Research Lab — reveals a remarkable convergence with the five that preceded it. The training science literature independently confirms what cognitive science (L0-001v2), motivation research (L1-002), assessment research (L1-003), instructional design research (L1-004), and competence formation research (L1-009) have all found from their respective vantage points:

- **Active engagement outperforms passive reception** (training's emphasis on practice; L1-004's ICAP framework; L1-003's testing effect).
- **Feedback must be fast, specific, and actionable** (training's tight feedback loops; L1-003's feedback design principles).
- **The environment shapes learning as much as the instruction** (training's psychological safety and error management culture; L1-009's environment-first principle; L1-002's SDT).
- **Assessment should inform, not just evaluate** (training's diagnostic assessment; L1-003's formative assessment; L1-002's undermining effect).
- **Competence is more than knowledge** (training's full-stack approach; L1-009's competence stack analysis; L1-008's knowledge-skills integration).
- **Time should serve learning, not the other way around** (training's mastery approach; Bloom's mastery learning; L1-003's criterion-referenced assessment).
- **Scaffolding should be faded progressively** (training's graduated responsibility; L1-004's expertise-adaptive model).
- **Errors are information, not failure** (training's error management; L1-004's productive failure; L1-009's error-as-information principle).

This convergence across independent research traditions — each with its own methods, journals, and academic cultures — is the strongest possible evidence that these are genuine principles of how humans learn. They are not artifacts of a particular discipline's biases or a particular researcher's agenda. They are what you find when you study learning seriously, wherever you look.

The training world's distinctive contribution to this convergence is not the principles themselves but the *demonstration that these principles actually work when implemented with real consequences.* Education researchers have identified these principles through controlled studies and meta-analyses. Training organizations have implemented them at scale, under conditions where failure is not tolerable, for decades. The evidence of their effectiveness comes not from effect sizes in journal articles but from the safety records of airlines, the competence of military forces, and the skill of surgeons.

Applied Pedagogy's challenge — and its opportunity — is to bring this convergence to bear on the design of educational experiences that develop the full competence stack, in contexts where the existential stakes are lower but the human stakes are no less real.

Dissertation complete.

BIBLIOGRAPHY

Aguinis, H., & Kraiger, K. (2009). Benefits of training and development for individuals and teams, organizations, and society. *Annual Review of Psychology*, 60, 451–474.

Baker, D. P., Day, R., & Salas, E. (2006). Teamwork as an essential component of high-reliability organizations. *Health Services Research*, 41(4p2), 1576–1598.

Baldwin, T. T., & Ford, J. K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology*, 41(1), 63–105.

Baldwin, T. T., Ford, J. K., & Blume, B. D. (2017). The state of transfer of training research: Moving toward more consumer-centric inquiry. *Human Resource Development Quarterly*, 28(1), 17–28.

Billett, S. (2001a). Knowing in practice: Re-conceptualising vocational expertise. *Learning and Instruction*, 11(6), 431–452.

Billett, S. (2001b). Learning through work: Workplace affordances and individual engagement. *Journal of Workplace Learning*, 13(5), 209–214.

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment*, 1(2), 1–12.

Blume, B. D., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: A meta-analytic review. *Journal of Management*, 36(4), 1065–1105.

Burke, C. S., Salas, E., Wilson-Donnelly, K., & Priest, H. (2004). How to turn a team of experts into an expert medical team. *BMJ Quality & Safety*, 13(suppl 1), i96–i104.

Burke, L. A., & Hutchins, H. M. (2007). Training transfer: An integrative literature review. *Human Resource Development Review*, 6(3), 263–296.

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

Eraut, M. (2004). Informal learning in the workplace. *Studies in Continuing Education*, 26(2), 247–273.

Gaba, D. M. (2004). The future vision of simulation in health care. *BMJ Quality & Safety*, 13(suppl 1), i2–i10.

Gagné, R. M. (1985). *The Conditions of Learning and Theory of Instruction* (4th ed.). New York: Holt, Rinehart and Winston.

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

Helmreich, R. L., Merritt, A. C., & Wilhelm, J. A. (1999). The evolution of crew resource management training in commercial aviation. *International Journal of Aviation Psychology*, 9(1), 19–32.

Issenberg, S. B., McGaghie, W. C., Petrusa, E. R., Lee Gordon, D., & Scalese, R. J. (2005). Features and uses of high-fidelity medical simulations that lead to effective learning: A BEME systematic review. *Medical Teacher*, 27(1), 10–28.

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*. New York: Hachette.

Keith, N., & Frese, M. (2008). Effectiveness of error management training: A meta-analysis. *Journal of Applied Psychology*, 93(1), 59–69.

Kulik, C.-L. C., Kulik, J. A., & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research*, 60(2), 265–299.

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

Marsick, V. J., & Watkins, K. E. (2001). Informal and incidental learning. *New Directions for Adult and Continuing Education*, 89, 25–34.

Marx, R. D. (1982). Relapse prevention for managerial training: A model for maintenance of behavior change. *Academy of Management Review*, 7(3), 433–441.

McGaghie, W. C., Issenberg, S. B., Cohen, E. R., Barsuk, J. H., & Wayne, D. B. (2011). Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. *Academic Medicine*, 86(6), 706–711.

Molenda, M. (2003). In search of the elusive ADDIE model. *Performance Improvement*, 42(5), 34–36.

Morrison, J. E., & Meliza, L. L. (1999). *Foundations of the After Action Review Process*. U.S. Army Research Institute Special Report 42.

Reason, J. (1990). *Human Error*. Cambridge: Cambridge University Press.

Salas, E., Burke, C. S., Bowers, C. A., & Wilson, K. A. (2001). Team training in the skies: Does crew resource management (CRM) training work? *Human Factors*, 43(4), 641–674.

Salas, E., DiazGranados, D., Klein, C., Burke, C. S., Stagl, K. C., Goodwin, G. F., & Halpin, S. M. (2008). Does team training improve team performance? A meta-analysis. *Human Factors*, 50(6), 903–933.

Salas, E., Tannenbaum, S. I., Kraiger, K., & Smith-Jentsch, K. A. (2012). The science of training and development in organizations: What matters in practice. *Psychological Science in the Public Interest*, 13(2), 74–101.

Salas, E., Wilson, K. A., Burke, C. S., & Wightman, D. C. (2006). Does crew resource management training work? An update, an extension, and some critical needs. *Human Factors*, 48(2), 392–412.

Sexton, J. B., Thomas, E. J., & Helmreich, R. L. (2000). Error, stress, and teamwork in medicine and aviation: Cross sectional surveys. *BMJ*, 320(7237), 745–749.

Tannenbaum, S. I., & Cerasoli, C. P. (2013). Do team and individual debriefs enhance performance? A meta-analysis. *Human Factors*, 55(1), 231–245.

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

Weaver, S. J., Dy, S. M., & Rosen, M. A. (2014). Team-training in healthcare: A narrative synthesis of the literature. *BMJ Quality & Safety*, 23(5), 359–372.

Weick, K. E., & Sutcliffe, K. M. (2007). *Managing the Unexpected: Resilient Performance in an Age of Uncertainty* (2nd ed.). San Francisco: Jossey-Bass.