Show Your Work: Functional Output as the Measure of Capability and Learning

Learning happens inside a person, where no one can see it directly; the only trustworthy evidence of it is what the person can do. A grade, a credit hour, and a passing score are each often used to certify that learning happened — while telling you nothing about whether it did. Based on a wide range of evidence drawn from education, engineering, and the other domains encompassed by learning engineering, we argue that professional competence and system capability must be measured by functional behavior.

Draft · August 2026 For an education conference on learning as demonstrated functional behavior
William Gray-Roncal · James Diamond

ON SCOPE · This paper argues a *position* about how professional competence should be evidenced. It is the ground a specialization like LENS stands on, not a description of that specialization: no course sequence, objective codes, or rubrics appear here. Its worked cases, its *What Fails / What Works* structure, and its evidence-typing discipline are drawn from *Capability Matters: A Casebook* (Gray-Roncal & Diamond, 2026) — a separately published casebook, complementary to but independent of any specialization — cited here as an evidence base, the way any published source is.

IN BRIEF

- **The claim.** Professional competence and system capability must be measured by *functional behavior* — a work sample scored against a real task — not inferred from a grade, a credit hour, or a test.
- **Why it bites.** Learning is invisible; behavior is its only trustworthy warrant, and cheap proxies routinely decouple from it — sometimes catastrophically (Watson for Oncology, the Epic Sepsis Model, Houston EVAAS).
- **The frame — KSAT.** A test certifies the *enablers* (Knowledge, Skills, Abilities); capability is the *Task*. Assess the task, and the enablers are entailed.
- **Human + system = capability.** Capability lives at the human-system interface — the operator's competence and the system's performance, combined — so a shortfall can be attributed to the system rather than blamed on the learner. (This is the turn past competency-based education, which assesses only the person.)
- **One frame, many domains.** The same standard — functional behavior, sampled in use — holds across grading, hiring, clinical AI, the checkride, and operational deployment. The interface it centers belongs to no single field; the cross-domain consistency is the contribution.
- **The two hard questions.** *Which* outcome to sample toward (push distal — toward results and the capacity to perform), and *what* evidence is enough (decision-grade: the fit between a decision's stakes and its evidence; the gold standard is "did it work?").
- **Fairer and feasible.** A demonstrated output is harder to inherit than a proxy (GEMS, Meyerhoff, CIRCUIT), and real problem-solving generates such outcomes in abundance.

A position offered to the learning-engineering community — IEEE ICICLE, LEBOK — not a standard.

ABSTRACT

A chemistry student scores 40% on the final; the curve turns 40 into a B; the transcript now certifies mastery. Nothing the student can *do* changed between the raw score and the letter. That gap — between the sign a credential records and the functional behavior it is taken to guarantee — is the subject of this paper. The position we call **demonstrated functional capability** holds that professional competence and system capability must be measured by functional behavior — a work sample scored against a real task, not inferred from a test, a grade, or a seat — and we argue it from a wide range of evidence across education, engineering, and the other domains learning engineering encompasses. Education has long known this — performance objectives, mastery learning, the OSCE, the checkride — but the proxy keeps winning because it is cheaper. We give the argument precise vocabulary through the **KSAT** decomposition of a job into Knowledge, Skills, Abilities, and Tasks, and take up the two questions it cannot dodge: *which outcome* the demonstration should sample toward (the criterion problem), and *what evidence* of it is enough (decision-grade evidence — the rigor a decision demands scales with its stakes, and the type of evidence caps the rigor a claim can reach). The one turn past even competency-based education is to locate capability at the *human-system interface* — **human + system = capability**, the operator's competence and the system's performance combined — so a gap can be attributed to the system rather than blamed on the learner. The contribution is the *interface* and its cross-domain *consistency*: competence assessment, system test-and-evaluation, and performance engineering are complementary

> views of one problem, and learning engineering sits where they meet — we translate among them, and claim the synthesis, not either half.

1 The forty that became a B

A student earns 40% on a chemistry final. The class curve turns 40 into a B. The transcript now says the student has command of chemistry. Nothing the student can *do* changed between the raw score and the letter — a norm-referenced curve measured the student's *rank in a cohort*, not their command of a reaction. The grade is a **sign**, and a peculiar one, because it was engineered to float free of performance: it certifies that the student did better than enough classmates, in a room that may have collectively understood very little.

The chemistry curve is not an anomaly; it is the mechanism, in miniature, of how credentials decouple from behavior. **Partial credit** rewards a derivation started confidently and finished wrong — the vessel does not care that the first two steps were right. The **credit hour** certifies *time in a seat*: the Carnegie unit was adopted in 1906 as an administrative convenience, and the foundation that named it has spent years since saying it was never a measure of learning. A course **taught to its test** raises the score without raising the capability — the two come apart precisely because the score became the target, which is Campbell's Law and Goodhart's Law stated as a lesson plan. Each is a proxy the real-world consequence never honors. The patient, the bridge, the cockpit, and the reaction vessel do not grade on a curve. These proxies persist for a reason deeper than cost, too: a credential is a *signal* in a market with imperfect information — what one side shows and the other trusts (Spence) — and a signal stays in use long after its tie to capability has frayed.

The distinction underneath all of them is old. In 1968 Wernimont and Campbell separated **signs** — traits, knowledge, and scores measured as *predictors* of performance — from **samples** — behavior elicited under conditions close enough to the job that the behavior *is* the performance. Miller's 1990 pyramid put it in a clinician's terms — *knows*, *knows how*, *shows how*, *does* — and observed that most assessment stops two rungs below the one that matters. Robert Mager had already written the corrective into instructional design in 1962: a real objective names *what the learner will be able to do*, under what conditions, to what criterion. This is not a claim that learning *is* behavior — learning is a durable internal change no one observes directly — but the recognition that behavior is the only trustworthy warrant we have for it. And because performance on any single occasion is an unreliable index of durable learning (Soderstrom and Bjork's learning-versus-performance distinction), a demonstration must reach for transfer and retention rather than settle for a one-time score — a caution §5 and §6 return to.

Education knows this, and keeps re-discovering it. Bloom's mastery learning refused to move on until you could perform. Competency-based programs award progress for demonstrated competencies rather than hours. Medicine largely replaced the oral exam with the **OSCE**, a circuit of stations where a candidate *does* the clinical task under observation — an advance whose ceiling we return to in §5. The trades kept the guild's **masterpiece**. Aviation never used anything else — the **checkride** is a practical test flown to a published standard. The through-line of every one is this conference's own premise, *learning demonstrated by functional behavior*, and the recurring failure of practice is that the demonstration is expensive and the proxy is cheap, so the proxy quietly wins.

And the pressure to close that gap is converging from several directions. Different constituencies reach the same conclusion from different starts: educators preparing students for tomorrow's problem-solving rather than yesterday's recall; colleges pressed to show graduates can apply book learning in the real world; employers moving from problem-solving puzzles to behavioral interviewing, because what a candidate has done predicts the job better than a whiteboard; and the military, aviation, and healthcare organizations that must reach operational readiness at speed and scale, where a credential not backed by demonstrated function is a risk they cannot carry. Different vocabularies, one requirement: measure capability and learning by functional behavior.

When the proxy wins in a classroom, a transcript overstates a student. When it wins in a high-consequence system, people are hurt. **Watson for Oncology** was sold worldwide on an advertised concordance with expert tumor boards as high as 96%, carried by the credibility of IBM and its MD Anderson partnership; leaked documents later showed it was trained on synthetic cases rather than real outcomes, produced unsafe recommendations, and was wound down — *the institutional credentials did not substitute for the validation.* **The Epic Sepsis Model** was the most widely deployed clinical AI in American medicine for four years before any independent evaluation; when one finally ran across 38,455 hospitalizations, its real AUROC was 0.63 against a vendor-reported 0.76–0.83, missing two-thirds of sepsis cases at the operating threshold. **Houston EVAAS** charged student test-score growth to individual teachers and drove terminations, until a 2017 federal ruling held that a score no one could inspect, replicate, or contest cannot support a consequential deprivation. Each produced a sign on schedule; none produced a demonstration anyone could check. They are the chemistry curve with the stakes turned up.

2 What “capability” is, and why it cannot sit inside a test

If demonstrated behavior is the standard, the object being demonstrated needs a definition, and the one this argument adopts is deliberately relational. **Capability is the interface between what a system requires of its human operators and the impact of that system.** It is not a property of a person. It is not a property of a machine. It is the *relation* between them, measured by whether the impact the system was fielded to deliver actually arrives. Put plainly, **human + system = capability**: the operator’s competence and the system’s performance combine at the interface, and what the world receives is their product — change either side, and the capability changes. This spans both halves the thesis names: a *person’s* competence and a *system’s* capability are the two sides of one interface (we develop the system side below). Two consequences follow, and both bear on assessment.

First, **capability is only observable in operation.** If capability is the operator-system interface under load, a measurement that removes the load — a proctored exam, a multiple-choice item, a recall task — has measured something adjacent to capability, not capability itself. This is a statement about the referent, not a complaint about test quality. There is no “the capability” waiting inside the person to be sampled by proxy, because half of it is on the other side of the interface.

Second, **a capability shortfall has more than one possible cause.** When the interface underperforms, the source may be system design, human development, organizational structure, or their interaction — and telling them apart is a competence in its own right, best called **gap attribution**. A test of the *person* structurally cannot perform it: it holds the system constant and varies only the person, so it returns “person” for every gap it can detect. The instrument’s design predetermines its finding. This is why the chemistry curve is not merely imprecise but *incapable* of the diagnosis that matters — a low class mean might be a cohort that did not study or a course that was not teachable, and the curve, by construction, cannot tell you which.

The system side of the interface is where most of the casebook’s evidence lives, and it has a mature vocabulary that says exactly what the human side says. Operational test and evaluation asks for *measures of effectiveness* and *performance* under realistic conditions, not a contractor demo; acquisition separates *demonstrated* from *paper* capability, and readiness levels ladder how far toward real use a capability has been shown; and where the stakes are safety, an *assurance case* — a structured, evidence-based argument that a system is adequately safe in use — is the systems-engineering twin of the validity argument the human side runs (§7). Each refuses the specification, or the demo, for the operational sample. A system certified by its advertised accuracy and never sampled in use is the chemistry curve one order of magnitude up — the Watson and Epic story (§1): capability *marketed* ahead of capability *validated*.

These two consequences are why a discipline serious about capability leans on a **body of demonstrated and failed cases** — the purpose a published casebook such as *Capability Matters* serves — rather than a body of doctrine to be examined on. The unit of the method is the iteration cycle — design, instrument, evaluate, refine — and the unit of its assessment is an artifact that shows one turn of that cycle actually completed.

3 The KSAT frame: tests certify the enablers; capability is the task

Workforce analysis has a standard vocabulary for decomposing a job, and it makes the layer confusion precise. A role is analyzed into **KSATs**:

K	S	A	T
Knowledge — what a person must know	**Skills** — what a person can do with practice	**Abilities** — enduring capacities the work draws on	**Tasks** — the observable units of work the role performs

The classic industrial-psychology triad was KSA (or KSAO, adding *Other characteristics*); modern occupational frameworks — the U.S. **O*NET** content model, the **NICE** Workforce Framework for Cybersecurity, and the DoD Cyber Workforce Framework, which uses the KSAT acronym explicitly — carry the fourth element, **Tasks**, as first-class, because the task statement is the thing an employer can observe and an auditor can verify.

The structural relationship is the whole point. **Knowledge, Skills, and Abilities are *enablers*; Tasks are *performances*.** A KSA is posited because it is believed necessary for some Task; it earns its place by the task it supports. This is why job-task analysis — the method underneath ISO/IEC 17024, the standard for bodies that certify persons — starts from the task inventory and *derives* the KSAs, not the other way around. The task is the criterion; the KSAs are hypotheses about what the task requires.

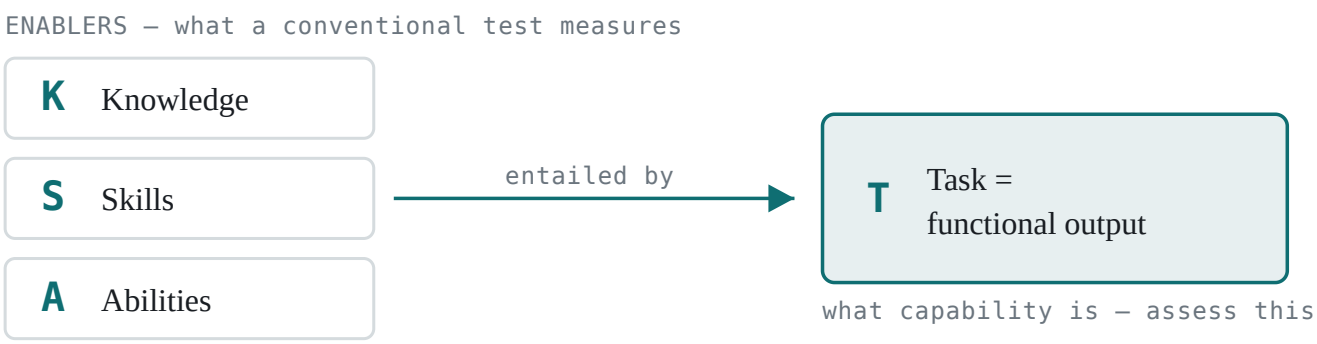

Figure 1. The KSAT layers. Knowledge, Skills, and Abilities are enablers; the Task is the performance. A conventional test samples the enabler layer and infers the task; demonstrated capability samples the task, and takes the enablers as entailed by the performance.

Human + **System** = **Capability**

the operator's competence and the system's performance, combined at the interface

By *functional output* we mean the Task made visible — a diagnosis rendered, a system fielded, a plan that survives contact — scored against what the work requires. Some capabilities (judgment, ethics, the exercise of agency) surface only indirectly in such outputs, which is one more reason to sample several rather than one.

A conventional test inverts the derivation. It assesses the K layer directly — the cheapest layer to write items for and the easiest to score reliably — and *infers* the Task. It says, in effect: this person holds the knowledge we believe the task requires, therefore we license the task. Every step of that inference is where the failures of §1 live. The knowledge may be necessary but not sufficient. The task may require an Ability the test never touched. The task may, in the operational setting, require *the system to hold up its half of the interface* — a condition entirely outside the person the test measured. There is even a coefficient for it: the U.S. Navy's Digital Tutor produced apprentices who *outscored* fleet technicians with an average 9.1 years of experience on a knowledge test, at an effect size of 4.30 — and the program still reported, unsmoothed, that knowledge accounts for only about 40 percent of variance on the practical exercise. A spectacular result at the K layer is a sign that leaves most of the sample unexplained.

The un-inverted posture is the one this position takes: **assess the Task, and let the demonstrated performance stand as evidence that the enabling KSAs are present.** A work sample that requires someone to attribute a capability gap, instrument an intervention, and produce decision-grade evidence under uncertainty need not separately test whether they *know* what decision-grade evidence is; the knowledge is entailed by the performance, the way finished joinery entails that its builder can read a level. It is the move the certification standards make when they require job-task analysis before a credential claim — which is why a program honest about this attaches no certification language until the bar for a personnel credential (a validated job-task analysis under ISO/IEC 17024 logic) is actually met.

There is a second, quieter reason KSAT matters. The DoD/NICE frameworks carried the fourth element because a *cross-functional* role — a leader, an integrator — is badly described by KSAs alone; what it contributes is legible only as tasks it performs and coordinates. That is the role this position most cares to certify: an **integrator** who can lead a mixed team, not a specialist who holds one KSA cluster deeply. Stating the *level* such a person reaches on each task — performs it independently, or briefs the specialist who holds it — is a task-and-role claim a knowledge test cannot make.

4 Different from education as usually practiced — and one turn past CBE

It would be easy to hear all of this as a familiar education argument: *authentic assessment beats multiple choice*. That is true, and not quite the one being made here. This position shares its floor with the best of the demonstrated-behavior tradition — Mager, mastery learning, the OSCE, competency-based education all assess what the learner can *do* — but it takes one further step, worth stating plainly to an audience that already believes the floor.

Competency-based education demonstrates the individual's behavior. Its object is the person; its success condition is that the person can now perform the task; its assessment, done well, is a work sample of that person. This is the top of Miller's pyramid, and a real advance over the grade and the credit hour.

A capability position demonstrates behavior at the interface. Its object is the operator-system relation; its success condition is that the system's intended impact arrives; and its question when impact does not arrive is *gap attribution* — a person to develop, a system to

redesign, an organization to restructure, or an interaction to re-engineer? On this frame a learner can be fully competent — every KSA demonstrated — and the capability still absent, because the system's half of the interface failed. And conversely, the right fix is often **not to teach the human at all**, but to change the design so the human need not hold the failing KSA. Two failure modes name this — a capability *designed out*, and a *human-system interface* that demands what no training can reliably supply — and neither is fixable by a better test of a person, however performance-based.

Thomas Gilbert's *Human Competence* (1978) gave this its foundational statement a decade before Miller: worthy performance is an accomplishment relative to its cost, and the largest, cheapest gains almost always come from the *environment* — information, instruments, incentives — rather than from the individual's repertory. Gilbert's Behavior Engineering Model is, read one way, a formal argument that even a good performance assessment can mislocate a problem if it looks only inside the person. Gap attribution is Gilbert operationalized and made assessable: someone must *demonstrate*, on a real case, that they can locate a performance gap correctly before proposing to close it — and a demonstration is the only assessment that can score whether they located it in the right place.

So this position stands *in* education — its foundation is the learning sciences (Dewey's learning by doing, situated cognition and communities of practice after Lave and Wenger, the cognitive-apprenticeship tradition of Collins, Brown, and Newman), its lineage the demonstrated-behavior tradition above — and holds this conference's premise harder than most practice does. What it adds is the engineer's move: capability is a property of an interface, not a possession of a person, so the demonstration has to be of the interface, and the diagnosis has to be free to send the fix somewhere other than the learner.

5 Exemplars: what the demonstration looks like, right and wrong

The everyday world is full of both halves of the distinction, and naming them is the fastest way to make the argument portable. The dividing line is exactly Wernimont and Campbell's: does the assessment ask about the person's attributes, or does it elicit a representative piece of the work?

● SIGNS — CERTIFY LITTLE ABOUT BEHAVIOR	● SAMPLES — CERTIFY BEHAVIOR DIRECTLY
× The curved chemistry grade (rank, not command)	✓ The road test that follows the written one
× Partial credit on a wrong derivation	✓ The clinical OSCE that replaced the viva
× The credit hour / seat-time (the Carnegie unit)	✓ The journeyman's masterpiece
× The driver's-license written test, taken alone	✓ The airline checkride, flown to standard
× A course drilled to its standardized exam	✓ The CPR renewal that re-tests the compressions

The **OSCE** is worth dwelling on, because it shows both the promise and the ceiling of demonstrated behavior. Replacing the oral exam with a circuit of stations where the candidate *does* the clinical task was a real advance: the functional output — took the history, examined the patient, communicated the finding — is directly observed, not inferred. But a demonstration can quietly become its own kind of proxy. A station is short, standardized, and often checklist-scored, and those properties reward **proximal** performance — the atomized behavior at *this* station in *these* five minutes — while under-sampling the **distal** capability it stands for: the integrated reasoning that transfers to an unstandardized patient, the outcome the encounter existed to produce. Checklist scoring can even invert the signal, rewarding the exhaustive novice over the expert who does less because they know which few things matter. So the *quality* of the demonstration — its fidelity to the real task, and whether it reaches distal reasoning or stops at proximal task-completion — decides whether the sample is a sample or a dressed-up sign. And transfer is hard and often narrow (Barnett and Ceci): a demonstration warrants a claim about *that* task, not about capability at large — one more reason a single sample is rarely enough.

The same line runs through the high-consequence record. *Capability Matters: A Casebook* (2026) collects it and organizes it in two halves — *What Fails* and *What Works — and the Frontier* — across seven high-consequence domains; the entries below are drawn from it, and worth reading in two columns. Most are *system* stories — the casebook lives on the system half of the interface — and each turns on whether a system's functional output arrived in use, the same test the human side sets.

SIGN PASSED, SAMPLE FAILED

Beyond Watson, Epic, and EVAAS (§1): **pulse oximetry across skin tones** is a validation-architecture failure that ran for thirty years — a bedside device cleared on an aggregate accuracy number that concealed a group-specific failure mode, systematically under-detecting hypoxemia in Black patients because validation was never demographically stratified. The aggregate sign did not merely proxy weakly; it *hid* the capability gap it was supposed to certify. **Three Mile Island** looks like a training failure, but the casebook attributes it elsewhere — a control room that reported *commands* rather than *states*, operators drilled only for dramatic ruptures rather than slow ambiguous cascades, and a near-identical Davis-Besse near-miss no channel ever carried to the fleet. A better-tested individual operator would have prevented none of the structural causes.

CAPABILITY DEMONSTRATED

TREWS / Sepsis Watch is the same delegation task as the Epic Sepsis Model and the opposite outcome — reduced mortality and organ failure in a prospective multi-site study — and the difference is precise: the capability deliverable is the alert design, the clinician role, and the outcome-grounded evidence loop, not the model. **Crew Resource Management and CAST** is the clearest evidence that capability is engineerable at the system level: Tenerife was not solvable by hiring better pilots, only by re-engineering the authority gradient in the cockpit — the paired cultural-plus-evidence intervention helped drive an 83% reduction in U.S. commercial-aviation fatality risk. And the **Navy Digital Tutor** is the anchor: its graduates beat nine-year veterans on the knowledge test *and* the program still refused to let the test stand for the capability.

The casebook's own failure taxonomy organizes cleanly against a proxy's monopoly. Of its six recurring modes, exactly one — a *training gap* — is what a better-taught or better-tested individual would have prevented. The other five are properties of designs, organizations, interfaces, and institutional memory: the environment side of Gilbert's model.

T Training Gap	**D** Designed Out	**N** Normalization of Deviance	**H** Human-System Interface
G Governance & Trust	**K** Knowledge & Institutional Memory		

An assessment that could only see the training-gap mode would be blind to five-sixths of the failure surface.

The same standard doubles as a fairer gate: scaffold students from non-traditional backgrounds so they can demonstrate functional capability *inside* the established system, then let the demonstration — not the pedigree — decide. Georgetown's **GEMS** postbaccalaureate program is the clean instance: complete it and beat the first-year medical-school (M1) average, and you earn a Georgetown interview — a gate opened by demonstrated performance at the level of enrolled students, not by a prior credential. UMBC's **Meyerhoff Scholars** has produced STEM PhDs from underrepresented backgrounds at rates that reset expectations, and **CIRCUIT** has moved students into research careers past structural barriers. Each measures itself, ultimately, by function delivered — an interview earned, a doctorate completed — which is at once the more valid signal and the more equitable one, because a demonstrated output is harder to inherit than a proxy. The gate is not automatically fair — a demonstration can also favor those with time, access, and coaching, and subjective scoring carries its own bias — which is why the scoring discipline of §§7–8 is part of the equity claim, not separate from it. To be successful we must identify and include the right people, and to judge that accurately, we have to see the work.

Functional outcomes, moreover, are not scarce — real problem-solving situations generate them continually, in the operational record, the program cohort, and the fielded system's results — and that is exactly the material against which learning hypotheses can be tested and performance iteratively improved. The abundance sits more naturally in real problem-solving than in traditional education paradigms, which are built to produce grades rather than outcomes; but the move is portable, and the classroom can borrow it by designing tasks whose functional outputs are real enough to study and learn from.

6 Which outcome? The criterion problem, and where the field has already worked it

"Demonstrate the capability" only pushes the question back a step: *which* capability, judged against *what* outcome? An assessment is no better than the criterion it samples toward, and choosing that criterion is a real and old problem — a position that argues for demonstration is incomplete until it says what the demonstration should be *of*. This is one of the most worked-over questions in the measurement, training-evaluation, and safety literatures. Four bodies of work bound the answer.

Taxonomies of the outcome — Bloom and after. Bloom's taxonomy (1956; revised by Anderson and Krathwohl, 2001) and its affective and psychomotor companions classify outcomes by *kind and cognitive level* — remember, understand, apply, analyze, evaluate, create.

Necessary discipline: “understands sepsis” and “manages a deteriorating patient” are different outcomes. But it fixes the *altitude* of an outcome, not whether it is the *right* outcome for the work — a perfectly specified create-level objective can still be the wrong thing to develop. (Bloom’s two-sigma effect is about the *magnitude* of outcome achievable — the lineage the Digital Tutor sits in.)

Levels of evaluation — Kirkpatrick, and its critics. Kirkpatrick’s four levels — reaction, learning, behavior, results — are a ladder of *how distal an outcome to hold an intervention to*: a knowledge test is level 2 (learning); demonstrated on-the-job behavior is level 3; organizational results are level 4 (Phillips adds return on investment as a fifth). Demonstrated functional capability lives at 3 and 4, not 2 — the whole quarrel with the grade and the certification. The critics (Holton, 1996; Alliger and Janak, 1989) sharpen the point: the levels are *not* automatically causally linked — a strong level-2 score need not yield level-3 behavior or level-4 results — which is exactly the reason to measure the distal outcome directly instead of inferring it from the proximal one.

Validity, and the criterion problem proper. Measurement has precise names for choosing the outcome badly: *criterion deficiency* (the measure misses part of the real outcome), *criterion contamination* (it captures what the outcome should exclude), and *criterion relevance* (the overlap that counts), from Thorndike (1949) through the standing “criterion problem” literature (Wallace’s *Criteria for what?*, 1965; Austin and Villanova, 1992). Messick’s unified validity (1989) and Kane’s argument-based validity (2013) give the modern frame: validity is not a property of an instrument but of the *interpretive argument* from a performance to a claim, and its two classic threats — construct underrepresentation and construct-irrelevant variance — are the formal names for the proximal/distal shortfall §5 found in the OSCE.

How applied fields pin the outcome — profession and safety. Two applied traditions answer “which outcome” concretely. In professional education the outcome is defined by the *work itself*: entrustable professional activities (ten Cate) name the units of practice a graduate should be trusted to perform unsupervised, and CanMEDS and the ACGME Milestones fix the developmental outcome by consensus grounded in job analysis — the same logic KSAT and ISO/IEC 17024 use (§3). And in safety science, where the ultimate outcome — the catastrophe — is rare and cannot ethically be waited for, the outcome is redefined from the *absence of harm* to the *capacity to perform under varying conditions*: Hollnagel’s Safety-II and resilience engineering, the high-reliability-organization literature (Weick and Sutcliffe), and the leading-versus-lagging-indicator distinction all say to measure the demonstrated ability to succeed, not only the tally of failures avoided. For a discipline built on high-consequence domains this is load-bearing — the appropriate outcome is a *capacity*, sampled before the accident, not a body count after it.

The convergent answer: there is no oracle that returns the one correct outcome, but there is a defensible procedure. Define the outcome from the real work (job-task analysis; entrustable activities). Operationalize it and *state what the operationalization leaves out* — designing the assessment before the content. Push it as far distal as the evidence can defensibly carry, toward Kirkpatrick’s results and Safety-II capacity rather than a proximal proxy. And hold it to a validity *argument* that names what is known, what is assumed, and what would change the decision. The capability-as-interface definition does one thing for free: by making the outcome the *system’s delivered impact*, it pins the criterion to results rather than learning — answering the criterion problem structurally rather than course by course.

7 What counts as enough? Decision-grade evidence is contextual, and the type of evidence sets the rigor

Choosing the right outcome (§6) is only half of the evidentiary problem. The other half is knowing when you have *enough* evidence that the outcome was met — and “enough” is not a fixed bar. This is what **decision-grade evidence** names, and the term is easy to misread as a synonym for “strong evidence.” It is not. Decision-grade is a *sufficiency judgment relative to a decision*, made under irreducible uncertainty: enough evidence, of the right kind, to justify *this* decision, at *these* stakes, now — while stating plainly what is known, what is assumed, and what would change the call. It is fit-for-purpose, not maximal.

STAKES & REVERSIBILITY RAISE THE BAR

A cheap, reversible choice can act on thin evidence; the cost of being wrong is a quick correction. An expensive, irreversible one — whose cost, in the high-consequence case, is paid by people who did not choose the risk — demands far more. Over-demanding rigor where a decision is cheap is its own failure: paralysis, or waiting for a trial the situation cannot afford.

EVIDENCE TYPE CAPS THE RIGOR

An investigation or peer-reviewed study carries more warrant than a program report, a practitioner account, a dissertation, or journalism. Type is a *ceiling* on the rigor a claim can honestly reach — so the honest move is to name the tier, not launder a weak source into a firm one.

The medical evidence hierarchy formalizes this — systematic review and randomized trial above cohort, case series, and expert opinion — and **GRADE** makes it a method: rate the *certainty* of the evidence first, then match the *strength* of the recommendation to it, never the reverse. The same discipline applies at the source: *Capability Matters* types each case by evidence source and flags the weaker tiers, letting “future validation ongoing” travel with a preprint or a news account rather than laundering it into something firmer. Weak evidence is not useless; it simply *caps* the rigor a claim built on it can honestly reach.

Put the two together and decision-grade is the **fit** between the rigor a decision demands and the rigor its evidence can supply. A journalism-tier account can be decision-grade for forming a hypothesis or halting a plainly harmful practice, not for fielding a clinical algorithm; a single-site program evaluation can be decision-grade for continuing that program, not for a national mandate. The record shows both errors: the **Epic Sepsis Model** failed the fit in the dangerous direction — national scale on vendor-internal evidence, no independent validation for years — while **TREWS** passed it, on prospective, multi-site, outcome-grounded evidence with the RCT-still-pending hedge carried rather than smoothed. Reading the *fit*, not the headline result, is the analytic skill: the same competence that attributes a gap (§2) has to say what grade of evidence a decision requires and whether the evidence in hand clears it.

The scaffolding programs of §5 — GEMS, Meyerhoff, CIRCUIT — make the discipline concrete. Each is a complex model of many steps and mentors, and instrumenting every step matters: it builds causal understanding, aims the next iteration, and gives learners the feedback that drives deliberate practice (Ericsson) — the same functional outputs that certify are what teach, so assessment *of* learning and *for* learning run on one instrument. But step-level evidence is not the gold standard. The gold standard is the distal question — *did it work?* — and a "yes" is decision-grade only when the controls and context carry the causal claim: compared to whom, selected how, in what setting, robust to which alternative. A program that produces doctorates or beats the M1 average has the gold-standard *signal*; whether it is *decision-grade* turns on whether the comparison ruled out that the students would have succeeded anyway. Instrument the steps to learn; judge the outcome to decide — at the grade the decision deserves.

This reflects back on demonstrated capability itself. A demonstration is evidence, and it too has a tier: a single work sample scored by the program that taught it is weaker warrant than the same capability shown across independent sites, scored by external assessors, and linked to real outcomes. So the strongest demonstrated-capability claim is one whose *own* evidence has been pushed up the hierarchy — externalized, replicated, outcome-coupled — and a position that sets a rigor bar for the systems it studies must hold its own evidence to it (§8).

---

8 What the position demands of an assessment

The argument would be idle if it did not constrain practice, so it is worth stating what any assessment that takes it seriously must do — as design commitments, not as any one program's machinery.

> **Aggregate many observations; never hang a high-stakes decision on one.** A single demonstration is unreliable — people perform inconsistently across tasks, and one rater is one rater — so triangulate multiple work samples, across contexts and assessors, into the decision. This is the logic of programmatic assessment (van der Vleuten and Schuwirth), and it is how reliability is bought back without retreating to the proxy.

> **Sample the task, criterion-referenced.** Score a representative piece of the real work against the standard the work sets, not against the cohort. A curve answers "better than whom?"; this must answer "good enough for what?"

> **Reach for the distal, and say where you stopped.** Prefer an integrated performance in a realistic context over an atomized station (§5), push the assessed outcome toward behavior and results rather than recall (§6), and state honestly where the demonstration stops short of transfer.

> **Report a level, not a pass.** State the proficiency reached — performs independently and produces evidence of it, versus recognizes it and can brief the specialist who holds it — following the INCOSE convention of proficiency levels per competency and the IBSTPI convention of competencies validated against practitioners.

> **Make the scoring inspectable, and grade your own evidence.** A rubric traceable to the task, a reproducible record, and an honest evidence tier on the demonstration itself keep a work sample from decaying back into a dressed-up sign (§7).

> **Instrument the program's own effect.** A discipline that tells others to produce decision-grade evidence of impact owes the same of itself: define entering and exiting capability as observable, measurable states, and evidence the delta the way it teaches everyone else to.

These commitments are where a concrete program turns the philosophy into curriculum, rubrics, and a capstone; ours is LENS, and those instruments live in our own documentation. The point here is only that the philosophy *has* teeth, and that they are these.

9 Relation to prior work: one frame across two literatures

This frame is new where it counts and old where it should be. Its novelty is the *interface* — capability as the operator-system relation — and the *consistency* with which one standard reads across domains that normally share no vocabulary. Its support is a set of mature, complementary traditions, each of which has already established a facet of it from its own side. The paper's work is explanatory: to translate among them and show they describe one thing. We invented neither half, and we say so.

On the **human** side, assessment science already treats a score as an *inference*, not a fact. Pellegrino, Chudowsky, and Glaser's assessment triangle (*Knowing What Students Know*) ties every claim about a learner to a model of competence, the observations that elicit it, and the interpretation that connects the two; evidence-centered design (Mislevy, Almond, and Lukas) builds assessments as exactly that evidentiary argument; entrustable professional activities and programmatic assessment (ten Cate; van der Vleuten and Schuwirth) certify by aggregated demonstration; validity theory (Messick, Kane) makes the performance-to-claim inference the object of scrutiny; and human performance technology — Gilbert's field — locates most performance gaps in the environment, not the person.

On the **system** side: operational test and evaluation, measures of effectiveness and performance, human systems integration, capability-based planning, and the safety- and assurance-case tradition already refuse the demo and the specification in favor of demonstrated function in use.

What none of them does is take the *interface itself* as the unit and carry it across domains. That is the move, and two things about it are new. First, **the interface as the object**: capability is neither in the person nor in the system but in their relation, so a shortfall can be located on either side — and no single tradition's tools, used alone, can make that call. Second, **consistency**: the same standard — functional behavior, sampled in use, held to decision-grade evidence — reads identically across a curved chemistry grade, a hiring decision, a clinical algorithm, a checkride, and an operational deployment. Each tradition above supports one facet of this; the contribution is to translate among these complementary capabilities and show that competence assessment, system test-and-evaluation, and performance engineering are one problem seen from three sides. That a single standard survives contact with domains this different is not a coincidence to explain away; it is evidence the frame is real. If the synthesis already has a name we have missed, we will adopt it — the frame matters more than the credit — but the recurring, expensive pattern the casebook documents says the interface is under-tended, not well mapped.

10 Limits, and what this does not claim

It is not anti-test. The Knowledge layer is real and worth assessing at its own grain; formative knowledge checks are owed to any learner. The claim is narrower: a test of the K layer cannot, by itself, *license the Task-layer claim*. Signs inform; samples certify. And for triage, very-large-scale screening, or well-validated low-stakes uses, a cheap sign is the right instrument — the objection is only to a sign *standing in for* a sample where the stakes demand the sample.

Demonstration is costly, and one sample is unreliable. Work-sample assessment trades reliability-per-dollar for validity: people perform inconsistently across tasks, assessors disagree, and setting a defensible "good enough" is hard. This is exactly why the cheaper proxy keeps winning — and why the answer is not one heroic capstone but many observations aggregated (§8), scored against inspectable rubrics. Demonstrated capability buys validity with effort, and the effort is the cost of the claim.

The sample can be gamed — and now cheaply faked. Make functional output the high-stakes target and Goodhart returns: coached answers, hollow-but-polished portfolios, teaching to the capstone — and generative AI makes fabricated "work" nearly free. The defense is the same as the attack: unpredictable, observed, in-context performance that is hard to fake — the checkride, not the take-home. As the fakes get cheaper, direct observation and authenticity matter more, not less.

Seeing the work is watching people. Observing real performance is data collection on people, and done carelessly it becomes surveillance — the casebook has its own school-surveillance case. The governance, consent, and fairness discipline this position demands of the systems it studies binds its own assessments too.

The distinctive competencies are the least settled. Gap attribution, outcome coupling, and capability at the human-system interface sit at the least-settled edge of the field's emerging consensus, where the evidence base is thinnest. That is an honest weakness: these are proposals, offered toward the IEEE ICICLE community's competency and body-of-knowledge work — LEBOK and the standards pathway it feeds — not asserted as settled.

Self-generated evidence is the standing risk. By §7's own standard, a program's evidence that its graduates can do these things is weak warrant if it is scored only by that program. The mitigation is non-negotiable and external: independent pilot sites, external assessors, and outcome coupling, so the evidence is not self-validating. A position that demands rigor of others fails on its own terms if it exempts itself.

Credentialing is deferred on purpose. A demonstrated-capability *assessment* is not yet a *credential*. A personnel credential requires a validated job-task analysis under ISO/IEC 17024 logic, and no adopted competency standard for learning engineering yet exists

to anchor one. Attaching certification language before that bar is met would be exactly the sign-for-sample substitution this paper argues against.

11 In plain language

A curve can turn a 40 into a B. A transcript can certify a degree. A benchmark can certify a model. None can tell you whether the learner, the clinician, or the operator can actually do the thing when the stakes are real — only a demonstration can: a real problem, an instrumented intervention, an artifact a stranger can check. But a demonstration is not automatically enough. It has to be *of the right outcome* — the capacity the work needs, not a proxy — and backed by *the right grade of evidence* for the weight of the decision it carries. Education has long known that the proof of learning is what a person can do, and keeps trading that proof for a cheaper sign. A serious professional discipline cannot make that trade — which means getting three things right at once: demonstrate the behavior, aim it at the outcome that matters, and hold the evidence to the standard the stakes deserve.

ACKNOWLEDGMENT

We are grateful to the community working to define learning engineering as a profession, whose efforts inform this paper: the **IEEE ICICLE** community — the International Consortium for Innovation and Collaboration in Learning Engineering — and its **Competencies, Credentials and Curriculum Special Interest Group**; the **Learning Engineering Body of Knowledge** (LEBOK), published openly at lebok.ai; and the competency-framework and standards efforts now underway. We have drawn on this work with appreciation. This paper offers one position; it does not speak for that community, and it presumes no particular competency framework or standard — none of which is yet settled. We share it as a contribution to an ongoing conversation, in the hope that it is useful.

Sources & further reading

COMPANION EVIDENCE BASE (INDEPENDENT PUBLISHED ARTIFACT)

— Gray-Roncal, W., & Diamond, J. (2026). ***Capability Matters: A Casebook*** (First Edition) — a separately published casebook of high-consequence successes and failures, complementary to but independent of any specialization. Source of the worked cases used throughout, the *What Fails / What Works — and the Frontier* organization, the six-mode failure taxonomy (§5), and the evidence-source typing discipline drawn on in §7.

PROGRAMS REFERENCED (§5)

— Georgetown University **GEMS** (Georgetown Experimental Medical Studies) postbaccalaureate program; UMBC **Meyerhoff Scholars** Program; **CIRCUIT** (see the casebook) — capability-scaffolding programs that gate on demonstrated function rather than prior credential.

FIELD CONTEXT — IEEE ICICLE AND THE STANDARDS EFFORT

— **IEEE ICICLE** — the International Consortium for Innovation and Collaboration in Learning Engineering, and its **Competencies, Credentials and Curriculum Special Interest Group**: the community defining learning engineering as a profession.

— **LEBOK** — *Learning Engineering Body of Knowledge*, published and openly available at lebok.ai (IEEE ICICLE): the field's knowledge map, and the reference this paper builds atop.

— The emerging learning-engineering **competency framework** and the **IEEE standards pathway** it feeds — the competency and proficiency layer above LEBOK, still being built; no adopted standard yet. This paper is offered toward that work.

DEMONSTRATED-BEHAVIOR LINEAGE IN EDUCATION

— Mager, R. F. (1962). *Preparing Instructional Objectives* — the performance objective (condition, behavior, criterion).

— Bloom, B. S. (1968). Learning for Mastery.

— Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: an integrative review. *Perspectives on Psychological Science*, 10(2) — current performance is an imperfect index of durable learning.

— Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine*, 65(9).

— Harden, R. M., et al. (1975). The Objective Structured Clinical Examination (OSCE). *BMJ*.

— Carnegie Foundation for the Advancement of Teaching — the Carnegie unit (credit hour) as time, not learning.

— Campbell's Law; Goodhart's Law — why a measure decays when it becomes a target.

LEARNING THEORY, TRANSFER, AND PRACTICE

— Dewey, J. (1938). *Experience and Education* — learning by doing.

— Lave, J., & Wenger, E. (1991). *Situated Learning* — communities of practice.

— Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship. Biggs, J. (1996). Constructive alignment.

— Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? — transfer is often narrow. Ericsson, K. A., et al. (1993). Deliberate practice.

WHY THE PROXY PERSISTS, AND HOW RELIABILITY IS BOUGHT BACK

— Spence, M. (1973). Job market signaling — credentials as signals under imperfect information.

— van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2005). Assessing professional competence: from methods to programmes — programmatic assessment aggregates many low-stakes observations into a high-stakes decision.

SYSTEM-SIDE EVALUATION AND THE ASSESSMENT-DESIGN FRAME (§§2, 9)

— Operational test and evaluation; measures of effectiveness and performance; capability-based planning — judging a fielded system by demonstrated function in use, not by specification or demo.

— Human Systems Integration (e.g., MANPRINT) — the human and the system as one design and evaluation object.

— Safety and assurance cases (e.g., Goal Structuring Notation) — a structured, evidence-based argument that a system is adequately safe in use; the systems-engineering twin of a validity argument.

— Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing What Students Know: The Science and Design of Educational Assessment* (National Research Council) — the assessment triangle: cognition, observation, interpretation. See also Pellegrino & Hilton (Eds., 2012), *Education for Life and Work*, on transferable knowledge and skills.

— Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A Brief Introduction to Evidence-Centered Design* — assessment as an evidentiary argument from work products to competence claims.

— Human Performance Technology (ISPI; Gilbert) — locating performance gaps in the environment, not only the person.

DEFINING THE OUTCOME — THE CRITERION PROBLEM (§6)

— Bloom (1956), *Taxonomy of Educational Objectives*; Anderson & Krathwohl (2001), the revision; Bloom (1984), the two-sigma problem.

— Kirkpatrick (1959/1994), the four levels; Phillips (ROI as a fifth); Alliger & Janak (1989); Holton (1996), *The flawed four-level evaluation model*.

— Thorndike (1949) — criterion deficiency / contamination / relevance; Wallace (1965), *Criteria for what?*; Austin & Villanova (1992).

— Messick (1989) — unified validity; construct underrepresentation & construct-irrelevant variance. Kane (2013) — argument-based validity.

— ten Cate (2005) — entrustable professional activities; CanMEDS; ACGME Milestones.

— Hollnagel (2014), *Safety-I and Safety-II*; resilience engineering; Weick & Sutcliffe, high-reliability organizations; leading vs lagging indicators.

EVIDENCE, SUFFICIENCY, AND RIGOR (§7)

— Levels-of-evidence hierarchies (e.g., Oxford CEBM); Sackett et al., evidence-based practice.

— **GRADE** — Grading of Recommendations Assessment, Development and Evaluation: certainty of evidence → strength of recommendation.

— Source typing and explicit evidence-tier flags as practiced in *Capability Matters* (investigation / peer-reviewed / program report / practitioner / dissertation / journalism; standing “future validation ongoing” qualifier) — naming the tier rather than laundering it.

ASSESSMENT AND COMPETENCE

— Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology*, 52(5).

— Gilbert, T. F. (1978). *Human Competence: Engineering Worthy Performance.*

— NICE Workforce Framework for Cybersecurity (NIST SP 800-181r1); DoD Cyber Workforce Framework (KSATs); O*NET Content Model.

— ISO/IEC 17024 (personnel certification; job-task analysis); INCOSE SE Competency Framework; IBSTPI competency standards.