ASSESSMENT, FEEDBACK, AND THE TESTING EFFECT

A Refined Review of Formative Assessment, Feedback Design, and Alternative Grading

Applied Pedagogy Research Lab

Guido Bartolucci, Principal Investigator

guido@appliedpedagogy.com

LAB.APPLIEDPEDAGOGY.COM

W2-003 · April 2026

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

CONTENTS

1

THE ASSESSMENT QUESTION, REFRAMED

Assessment sits at the intersection of three pressures that rarely align. The cognitive pressure is clear: retrieval practice and formative feedback are among the most robust learning levers the science of learning has identified (Roediger & Karpicke, 2006; Yang et al., 2021)○. The motivational pressure pulls in the opposite direction: controlling assessment — grades, rankings, high-stakes testing — undermines the intrinsic motivation that makes students want to learn in the first place (Deci, Koestner & Ryan, 1999; Butler, 1988)○. The institutional pressure complicates both: schools and accrediting bodies demand comparable, documentable assessment data, and neither the cognitive-science prescription ("test frequently, at low stakes") nor the motivational prescription ("eliminate evaluative control") is easily reconciled with what institutions require.

The v1 review named this three-way tension clearly. It documented the testing effect, synthesized the feedback literature through the Hattie-Timperley (2007) framework, engaged the assessment-motivation problem through the lens of self-determination theory, and offered a practical model organized around seven principles. What v1 could not do — because it relied primarily on training knowledge and did not engage the primary texts at depth — was resolve the tension it identified. It named the paradox and offered principles. It did not narrow the paradox or specify what a concrete assessment system would look like.

This review attempts to go further. It engages the primary texts that v1 cited at summary depth — Black and Wiliam's (1998) review, Wiliam's (2011) *Embedded Formative Assessment*, Hattie and Timperley (2007), Kluger and DeNisi (1996), Butler (1988), Kohn's (1993) *Punished by Rewards*, and Blum's (2020) *Ungrading* — through the PI's book-summarization pipeline and open-access sources. It extends the investigation into four territories v1 did not enter: feedback in ill-structured domains (writing, design, clinical reasoning, and the arts), AI-assisted feedback and its emerging risks, cross-cultural variation in assessment effects, and the practitioner literatures on standards-based grading and ungrading. It connects to three other W2 reviews: W2-008's reframing of self-regulation as environmentally conditioned rather than trainable, W2-009's treatment of feedback for judgment at Layer 3 of the competence stack, and W2-001's primary-literature engagement with retrieval practice mechanisms.

The question this review addresses is not "does assessment work?" but something harder: *How can assessment and feedback be designed to maximize learning without destroying the motivation to learn — across well- and ill-structured domains, across short and long time horizons, and at scales where real teachers work?*

The binary formative-versus-summative distinction that organized v1's argument is a useful entry point but an inadequate frame. As Wiliam (2011) argues, the distinction is functional, not instrumental — the same assessment event can operate formatively or summatively depending on how its evidence is used (§8.0.4)●. The more productive frame, and the one this review adopts, organizes assessment by the question it answers: Does this assessment event generate evidence that someone — the teacher, the learner, or a peer — can use to make the next learning step better? If so, it functions formatively regardless of its form. Does it generate a grade, a rank, or a credential? If so, it functions summatively regardless of its intent. The design challenge is to maximize the former and minimize the damage the latter does to the conditions that make the former effective.

W2-008's normative framework clarifies what assessment is *for*. Assessment in the Applied Pedagogy frame is not a measurement problem — it is a developmental problem. The compe-

tence stack (knowledge, skill, judgment, metacognition, character) described in the lab's outcome specification requires assessment practices that operate differently at each layer. Knowledge and skill (Layers 1–2) can be assessed with relatively conventional tools: retrieval practice, diagnostic questioning, skill demonstrations. Judgment and metacognition (Layers 3–4) require the kinds of feedback that the standard assessment literature barely engages — feedback for calibration under uncertainty, for the development of evaluative capacity, for the cultivation of what Sadler (2010) calls "guild knowledge." Character (Layer 5) may not be directly assessable at all, but the assessment climate shapes it powerfully — W2-008's finding that self-regulation is environmentally conditioned implies that assessment design is itself a character-formation mechanism, whether or not it intends to be.

2

THE FORMATIVE-ASSESSMENT CASE

2.1 BLACK AND WILIAM (1998): THE LANDMARK AND ITS DEFLATION

The modern formative-assessment movement begins with Black and Wiliam's (1998) review, "Assessment and Classroom Learning," published in *Assessment in Education*. The paper has accumulated over 7,400 citations and a field-weighted citation impact of 158.7. Black and Wiliam reviewed approximately 250 studies and concluded that formative assessment — assessment designed to provide feedback that moves learning forward, rather than merely to assign grades — produces substantial learning gains. The effect sizes they reported ranged from 0.4 to 0.7 (Abstract-verified).

These numbers have been consequential. They launched a global movement in educational policy, funded professional development programs across the English-speaking world, and established formative assessment as a premier intervention. They are also, by the best available meta-analytic evidence, overstated.

Kingston and Nash (2011), in a more methodologically rigorous meta-analysis, screened over 300 formative-assessment studies. Only 13 provided sufficient data for meta-analysis — a striking finding in itself, suggesting that the vast majority of formative-assessment studies do not meet basic methodological standards for causal inference. The weighted mean effect across those 13 studies was d = 0.20, roughly one-third of the lower bound of Black and Wiliam's range. Science yielded d = 0.09; mathematics and English/language arts were somewhat higher. The strongest effects appeared in studies where professional development supported implementation, yielding d = 0.30 — still half of Black and Wiliam's lower bound (Abstract-verified, W2111762371).

More recent meta-analyses have confirmed Kingston and Nash's deflation rather than Black and Wiliam's original claims. Lee et al. (2020), restricting to U.S. K-12 studies with control conditions, found an overall mean of d = 0.29, with student-initiated self-assessment showing the highest effects at d = 0.61 (Abstract-verified, W3010246167). Sortwell et al. (2024), in an umbrella review of 13 formative-assessment meta-analyses, assigned GRADE certainty ratings and found 9 of 13 rated "very low certainty" — meaning that the true effect could be substantially different from the estimated one. The range they report — "trivial to large" — describes real heterogeneity, not a settled estimate (Abstract-verified, W4402374697, FWCI 86.82).

The honest claim, grounded in the strongest meta-analytic evidence available, is this: formative assessment in well-controlled K-12 studies produces effects in the range of d = 0.20–0.30. These effects are real and educationally meaningful — a d of 0.25 represents roughly two to three months of additional learning per year. But they are a different order of magnitude from the d = 0.4–0.7 that launched the movement, and the certainty ratings on even these more modest estimates are low.

Why the discrepancy? Part of the answer is methodological: Black and Wiliam's 1998 paper was a configurative review, not a formal meta-analysis. They were explicit about this: "We did consider at this point conducting a formal meta-analysis of the studies we had identified, but we quickly realized that with such a diverse range of studies, meta-analysis would simply not be appropriate" (Wiliam, 2011, §7.0.12)[•]. The effect-size range appeared in a separate practitioner publication as

an illustrative estimate, not as the output of a systematic aggregation. Subsequent researchers and policymakers treated it as a definitive finding.

But the deeper answer lies in what Black and Wiliam themselves recognized in their 2018 reassessment: "the relationship between assessment and instruction is far more complex than their earlier work had suggested" (Abstract-verified, W2770208442, FWCI 151.71). Formative assessment, they acknowledged, had been extracted from its pedagogical context and treated as a standalone technique. The 1998 framework was "incomplete." When formative assessment is implemented as a bolt-on — a technique added to existing practice without changing the teacher's pedagogical stance — the effects are modest. When it functions as Wiliam (2011) describes — as a wholesale transformation of the teacher's role from transmitter to diagnostician — the effects are larger. The deflation in meta-analytic effect sizes reflects the fact that most implementations are bolt-ons.

2.2 WILIAM'S FIVE STRATEGIES

Wiliam's (2011) *Embedded Formative Assessment* develops the theoretical case into five operational strategies, derived from a 3×3 matrix crossing three processes (finding out where learners are, establishing where they are going, determining how to get there) with three roles (teacher, peer, learner) (§9.0.2)●. The strategies are: (1) clarifying, sharing, and understanding learning intentions and success criteria; (2) engineering effective discussions, activities, and tasks that elicit evidence of learning; (3) providing feedback that moves learning forward; (4) activating learners as instructional resources for one another; and (5) activating learners as owners of their own learning.

Strategy 3, on feedback, is the most extensively developed in the book and connects directly to the feedback literature reviewed in Section III. Strategies 4 and 5 — peer feedback and self-assessment — are the most underused in practice and potentially the most important for the assessment-motivation tension. Wiliam reports that students engaged in structured self-assessment in a Portuguese mathematics study compressed 38 weeks of typical learning into 20 weeks — though he acknowledges this rests on a single unreplicated study (§29.0.5–29.0.6)◑. The mechanism is consistent with the testing-effect literature: self-testing (retrieval practice) is among the two highest-utility study strategies across 400+ studies, and Wiliam specifies that "self-tests should be ungraded and unrecorded — the benefit comes from retrieval practice, not formal evaluation" (§30.0.17)◑.

2.3 WHY FORMATIVE ASSESSMENT IS HARD TO SUSTAIN

The implementation gap is the formative-assessment literature's defining practical problem. Schildkamp et al. (2020) identify three structural prerequisites for sustained formative assessment: teacher knowledge and skills (diagnostic capacity, content knowledge sufficient to interpret student thinking), psychological factors (willingness to change, tolerance of ambiguity), and social factors (collegial support, institutional permission to deviate from standard practice). Most implementations fail because one or more of these prerequisites is absent (Abstract-verified, W3036375580).

Wiliam's own implementation advice is deliberately conservative: "When teachers try to change more than two or three things about their teaching at the same time, their teaching typically deteriorates, and they go back to doing what they were doing before" (§33.0.20)●. His recommended structural form for sustaining change — teacher learning communities meeting monthly to discuss implementation — operates on an 18-month time horizon for embedding two or three

techniques. This is a realistic timeline. It is also radically slower than the policy timelines that typically accompany formative-assessment mandates.

The institutional environment often works directly against formative assessment. When schools are evaluated primarily on summative test scores, the incentive structure pushes toward teaching to the test — a form of summative assessment that crowds out the formative practices that would actually improve the learning being tested. This is the testing paradox that Ryan and Weinstein (2009) identified from an SDT perspective: high-stakes accountability systems create controlling environments that undermine both the quality of teaching and the quality of learning (Abstract-verified). The paradox is structural, not individual: it is not that teachers lack the will to implement formative assessment, but that the institutional reward structure makes formative assessment an unrewarded activity.

3 THE FEEDBACK LITERATURE

3.1 THE SURPRISING FRAGILITY OF FEEDBACK

The common assumption that feedback improves learning is mostly wrong — or at least dangerously oversimplified. Kluger and DeNisi (1996) conducted the most comprehensive meta-analysis of feedback interventions to that date, analyzing 607 effect sizes from 131 studies spanning nearly a century of research. Of the approximately 3,000 studies they screened, only 131 — about 4% — met basic scientific-quality criteria. Their headline finding was startling: feedback interventions improved performance on average (d = 0.41), but decreased performance in over one-third of cases. In Wiliam's formulation: "This is one of the most counterintuitive results in all of psychology. In each of the studies, feedback was intended to improve performance, but in almost two out of every five carefully conducted studies, the participants would have done better if the feedback had simply not been given" (2011, §22.0.50)•.

This 38% finding should be the starting point of any honest discussion of feedback, not a footnote. It means that the intuitive practice of "giving students feedback" is not reliably beneficial. Whether feedback helps or hurts depends on what kind of feedback, about what, to whom, and in what context. Kluger and DeNisi proposed Feedback Intervention Theory (FIT) to explain the pattern. The theory distinguishes three levels at which feedback can direct attention: the task-learning level, the task-motivation level (meta-task processes), and the self level. The key prediction is that feedback becomes less effective — and potentially harmful — as attention moves from task-learning toward the self. When feedback tells the learner "your calculation in step three contains an error," attention stays on the task. When feedback tells the learner "you're really struggling with math" — or, more subtly, when a grade implies "you are a B student" — attention shifts to the self, triggering ego-protective processes that interfere with learning (Abstract-verified, W2030441548).

3.2 THE HATTIE-TIMPERLEY FRAMEWORK

Hattie and Timperley (2007) built on Kluger and DeNisi to develop the most widely cited framework for understanding feedback in education. Their paper, with over 11,500 citations and FWCI of 482.1, proposes that effective feedback answers three questions: Where am I going? (feed-up), How am I going? (feed-back), and Where to next? (feed-forward). They distinguish four feedback levels (Abstract-verified, W2560140854):

Task-level feedback addresses correctness, accuracy, or completeness. It is the most common form in classrooms and the most appropriate for novices encountering new material. It is also the least transferable — correcting an error on this problem does not help the student approach the next problem differently.

Process-level feedback addresses the strategies and error- detection approaches the learner used. It is more transferable than task-level feedback because the strategies apply across tasks. In Wiliam's metaphor, it is the difference between "you're not dropping your pitching shoulder enough to deliver the pitch from below the knee" (actionable, process-level) and "you need to get your ERA down" (accurate but useless) (2011, §23.0.5–23.0.7)•.

Self-regulation feedback develops the learner's capacity to monitor and direct their own learning — internal quality control. This level is most powerful for learners who already have sufficient domain knowledge to engage in meaningful self-monitoring.

Self-level feedback addresses the person rather than the work — praise, encouragement, or criticism directed at the self. This level is the least effective and potentially the most harmful. Even well-intentioned praise, when directed at the person ("You're so smart") rather than the process, shifts attention from task to self and creates dependence on external validation.

3.3 THE WISNIEWSKI UPDATE

Wisniewski, Zierer, and Hattie (2020) conducted an updated meta-analysis that provided quantitative support for the framework. Their analysis found an overall effect of d = 0.48 — substantially lower than the d = 0.79 from Hattie's Visible Learning synthesis, which was methodologically problematic (fixed-effects model, failure to remove duplicate studies across meta-analyses). The variance was enormous: some feedback conditions produced effects above d = 1.0, while others produced negative effects. The most important moderator was not timing, frequency, or delivery mode — it was content. What the feedback said mattered more than how or when it was delivered. High-information feedback (approximating process and self-regulation levels) yielded d = 0.99. Reinforcement and punishment yielded d = 0.24. Praise had minimal or negative effects (Abstract-verified, Wisniewski et al., 2020).

This finding has a critical practical implication. Much of the conversation about feedback in education focuses on logistics — how quickly to return papers, how often to give quizzes, whether to use written or verbal feedback. These considerations are not irrelevant, but they are secondary. The fundamental question is what the feedback communicates. Feedback that directs attention to the task and how to improve it is the most consistently beneficial. Feedback that directs attention to the self — even positive self-level feedback — is the least beneficial and sometimes harmful.

3.4 THE BUTLER FINDING: GRADES CANCEL COMMENTS

Butler (1988) demonstrated with unusual clarity what happens when feedback and evaluation collide. Students in three conditions — comments only, grades only, and comments plus grades — showed strikingly divergent outcomes. Students who received comments only improved by approximately 30 percentage points. Students who received grades only made no progress. And students who received both grades and comments performed identically to students who received grades only. As Wiliam (2011) puts it: "giving grades alongside the comments completely washed out the beneficial effects of the comments; students who got high grades didn't need to read the comments, and students who got low scores didn't want to" (§22.0.8)•.

Butler's study involved Israeli fifth and sixth graders on researcher-designed tasks over two sessions — a limited scope. No direct replication has been published. The high-achiever exception should be noted: high achievers who received grades maintained high interest when further grades were anticipated (Abstract-verified, W2162070030). But the core finding — that the presence of a grade negates the informational value of accompanying comments — is consistent with Kluger and DeNisi's FIT mechanism and with the subsequent evidence on how grades function as extrinsic rewards. When students see a grade, the grade becomes the message.

3.5 FEEDBACK LITERACY

Carless and Boud (2018) introduced the concept of feedback literacy — the capacity to make productive use of feedback. Their paper, with over 1,800 citations and FWCI of 363.4, identified four competencies: appreciating feedback (viewing it as information rather than judgment), making judgments (the ability to evaluate one's own work against standards), managing affect (processing critical feedback without defensiveness), and taking action (converting feedback information into specific behavioral changes) (Abstract-verified).

The feedback literacy framework reframes the feedback problem. The traditional question — "How do we give better feedback?" — turns out to be half the question. The other half is: "How do we develop students' capacity to receive, interpret, and act on feedback?" Simply providing better feedback is insufficient if students lack the feedback literacy to use it. Winstone et al. (2017) identify structural recipience barriers — if there is no opportunity to act on feedback (because the assignment is a one-shot submission with no revision cycle), behavioral recipience is impossible regardless of the feedback's quality (Abstract-verified). Carless and Boud's framework implies that developing feedback literacy requires explicit instruction, structured practice in evaluating work against criteria, and — crucially — revision opportunities that make feedback actionable.

3.6 FEEDBACK TIMING: MORE COMPLICATED THAN IT SEEMS

The question of when to give feedback resists a simple answer. For factual error correction, immediate feedback is superior — it prevents the wrong answer from consolidating in memory (Butler & Roediger, 2008)○. But for conceptual learning, the picture is different. Wiliam (2011) reports the Mullet et al. (2014) engineering study: students receiving delayed feedback averaged 93% on end-of-course exams versus 84% for immediate feedback. Ninety percent of students reported preferring immediate feedback, and 79% believed they benefited more from it — both contrary to the actual outcomes (§22.0.99–22.0.103)●. Wiliam's conclusion: "in general, there is an inverse relationship between the quality of performance on a task and the amount of learning that occurs as a result of completing that task" (§22.0.102)●.

This finding is consistent with Bjork's desirable-difficulties framework and with the broader testing-effect literature: conditions that feel harder during learning often produce more durable outcomes. The practical implication is not "always delay feedback" but rather that timing must be matched to the learning target. Immediate corrective feedback is appropriate for procedural errors. Delayed feedback may be superior for conceptual development, because the delay preserves the productive struggle that deeper encoding requires. For ill-structured tasks — where the evaluative criteria are themselves being learned — the timing question is secondary to whether feedback is informational or controlling.

4 THE TESTING EFFECT IN CLASSROOMS

4.1 THE COGNITIVE FOUNDATION

The testing effect — the finding that practicing retrieval strengthens memory more than additional study time — is one of the most robust findings in cognitive psychology. Roediger and Karpicke (2006) provided the foundational demonstration: students who studied a passage and then took a practice test retained significantly more after one week than students who studied the passage twice. Crucially, students predicted the opposite. The testing effect is not just real; it is counterintuitive. W2-001's review engages the cognitive mechanism at depth — effortful retrieval strengthens memory traces through elaborative processing and the generation of additional retrieval cues (W2-001, §III). This review cites W2-001's treatment rather than duplicating it and focuses on the classroom-translation and assessment-design questions that are specific to the assessment domain.

4.2 CLASSROOM TRANSLATION

Yang, Luo, Vadillo, Yu, and Shanks (2021), in a systematic review and meta-analysis focused specifically on the translation of the testing effect from laboratory to classroom, found an overall effect of $g = 0.49$ in classroom studies — slightly smaller than laboratory effects but still substantial. The effect held across subject areas (STEM, social sciences, humanities), educational levels (elementary through university), and assessment formats (Abstract-verified). This finding provides the critical ecological-validity bridge: the testing effect is not a laboratory curiosity — it replicates in real classrooms.

Four boundary conditions from the meta-analytic evidence must qualify the headline number:

First, material complexity matters. Van Gog and Sweller (2015) argue that the testing effect decreases and may disappear as element interactivity increases — for multi-step problems requiring integration of multiple concepts, the benefit is smaller than for factual recall (Abstract-verified, W2128915986). Yang et al. confirm this in classroom data: factual and procedural learning show effects of $g \approx 0.55–0.65$, while conceptual understanding shows $g \approx 0.40$. For far transfer, the evidence is thinner and more variable. Adesope, Trevisan, and Sundararajan (2017) found effects for application-level outcomes, but these were smaller than effects for recall (Abstract-verified).

Second, testing without feedback is substantially weaker than testing with corrective feedback. Roediger and Karpicke (2006) themselves acknowledge that tests without feedback can entrench errors — if a student retrieves the wrong answer, the retrieval can consolidate the error. Butler and Roediger (2008) showed that corrective feedback eliminates the misinformation effect of multiple-choice testing, where plausible-but-wrong distractors can introduce errors (Abstract-verified, W2143805697). The combination of retrieval attempt plus corrective feedback is more powerful than either alone.

Third, the educational level moderates the effect. Yang et al. find stronger effects in higher education ($g \approx 0.55–0.60$) than in K-12 ($g \approx 0.40–0.45$), with elementary effects smaller and more variable. This likely reflects the interaction between the testing effect and learner self-regulation capacity: younger learners may not be able to monitor their own retrieval quality as effectively as older students (Training-derived, consistent with the developmental literature).

Fourth — and most important for assessment design — the stakes matter. The testing effect is a finding about *low-stakes retrieval practice*, not about *high-stakes summative testing*. Yang et al. explicitly distinguish the two: low-stakes retrieval practice functions as a learning tool; high-stakes testing functions as an evaluation mechanism. The cognitive benefit does not transfer to the motivational consequences of testing-as-accountability. Wiliam (2011) is explicit: the benefit of retrieval practice comes from the retrieval, not from formal evaluation, and self-tests should be ungraded and unrecorded (§30.0.17)[1]. When retrieval practice is graded, it shifts from a learning mechanism to an evaluation mechanism — and the Butler (1988) finding applies: the grade cancels the learning benefit.

4.3 THE INTEGRATION: TESTING EFFECT MEETS FORMATIVE ASSESSMENT

The testing effect and formative assessment literatures developed in separate intellectual traditions — cognitive psychology and educational measurement, respectively — but they converge on the same practical prescription. A brief, low-stakes quiz at the beginning of class simultaneously exercises retrieval (strengthening memory), generates formative evidence (showing the teacher and students what has and has not been learned), and — if kept ungraded — preserves the informational assessment environment that SDT predicts will support rather than undermine motivation.

This convergence means the evidence base for frequent low-stakes assessment is not just the formative-assessment literature or the testing-effect literature. It is both. The cognitive mechanism (retrieval strengthens memory), the instructional mechanism (evidence enables adaptive teaching), and the motivational mechanism (low-stakes assessment is informational rather than controlling) all point in the same direction.

5

ALTERNATIVE ASSESSMENT SYSTEMS

5.1 THE GRADING PROBLEM

The evidence reviewed in Sections II–IV creates a practical problem. The most effective assessment practices — frequent low-stakes retrieval, comments without grades, formative assessment that generates evidence for adaptive teaching — are the practices most incompatible with conventional grading. The most common assessment practice — letter grades accompanied by evaluative comments — is the practice Butler (1988) showed is equivalent to grades alone. Something must give.

The standard response in the assessment literature is to propose alternative grading systems that preserve the accountability function while reducing the motivational damage. Three major alternatives have been developed: standards-based grading, ungrading, and competency-based assessment. Each has a different evidence base, a different institutional fit, and a different set of unresolved problems.

5.2 STANDARDS-BASED GRADING

Standards-based grading (SBG) replaces traditional letter grades and percentage scores with ratings against specific learning standards. Instead of a single "B" on a unit, a student receives separate ratings on each standard or competency — "proficient" on algebraic reasoning, "developing" on geometric proof. This approach is more informational than traditional grading because it disaggregates performance into actionable dimensions.

Marzano's (2000) *Transforming Classroom Grading* provides the foundational practitioner framework. His central diagnostic is devastating for conventional grading: A-students in high-poverty schools score at C-/D+ levels on standardized assessments (§0.9.10)●. This referencing problem — the same grade means different things in different contexts — makes traditional grades unreliable as measures of actual learning. Marzano proposes a four-point rubric system (beginning, developing, proficient, advanced) with topic-specific descriptors and criterion-referencing. His rubric structure is task-level in the Hattie-Timperley sense — it tells the student where they stand relative to a defined criterion — but the descriptors ("complete and detailed understanding" versus "significant gaps") are often too imprecise to constitute actionable process-level feedback.

Schimmer's (2016) *Grading from the Inside Out* extends the SBG practitioner framework with an emphasis on the teacher's mindset shift. His contribution is practical rather than empirical: how to actually implement SBG given the institutional barriers of existing grading policies, parent expectations, and student resistance. He provides no original outcome data (Pipeline/Practitioner). Wormeli's *Fair Isn't Always Equal* (2018) addresses the middle-school implementation of SBG with substantial practical detail. O'Connor's *A Repair Kit for Grading* (2010) focuses on eliminating the most damaging traditional grading practices — averaging, giving zeros, combining academic and non-academic factors — as incremental reforms that do not require wholesale SBG adoption.

The evidence problem with SBG is severe. Link and Guskey (2022), in the most direct academic review of SBG evidence, found that SBG effectiveness cannot be evaluated because there is no stable consensus definition of what SBG is. Three definitional criteria — criterion-referenced grading,

reporting by standard, and separating academic from non-academic factors — are inconsistently applied across implementations. "Uncertainty, confusion, frustration, and resistance" are causing abandonment in some districts (Abstract-verified, W4288051987). Whether SBG improves learning, motivation, or achievement is not established by the academic literature.

Marzano's evidence base is McREL-internal and illustrative — the examples are vivid but the evaluation is self-published (Verified via PI summary). The grey literature on district-scale implementation — school district reports, state policy evaluations — was not systematically searched and represents a gap in this review's coverage.

5.3 THE UNGRADING MOVEMENT

Ungrading — the practice of removing or radically minimizing grades, typically replacing them with self-assessment, narrative feedback, or contract grading — has consolidated since Blum's (2020) edited volume. The volume is the anchor text, and engaging it at primary-text depth reveals both its strengths and its limitations more clearly than v1's summary treatment could.

Blum's core thesis is that grades are not a neutral measurement technology but a historically contingent institutional invention — originating from William Farish's 1792 Cambridge innovation and the Mount Holyoke 1897 letter-grade system — that actively undermines intrinsic motivation, encourages surface-level performance, and serves sorting functions rather than pedagogical ones (Verified via PI summary). The theoretical case draws on Butler (1988), Deci and Ryan's SDT, and Kohn's *Punished by Rewards* (1993). The empirical case is almost entirely practitioner testimony.

This distinction — between a strong theoretical case and a weak empirical base — is the honest assessment of the ungrading literature. Zero RCTs, zero quasi-experimental studies, zero blinded assessments appear in the Blum volume. Every contributor who adopted ungrading reports net positive outcomes — a substantial publication/selection bias that the book partially acknowledges. Blum herself states the control- condition problem directly: "rigorous comparative research on ungrading outcomes is elusive because practitioners refuse to revert to traditional grading as a control condition" (§2.124.2)●.

The implementation variants in the volume illustrate both the range and the limitations. Blum's portfolio-plus-reflection model involves student-proposed final grades after semester- long portfolio review. Inoue's labor-based grading contracts determine grades by documented effort rather than assessed quality. Riesbeck's critique-driven gradeless course at Northwestern has sustained iterative submission-feedback- revision cycles for over two decades in computer science — the most technically developed and longest-running implementation, but in a well-structured domain where "acceptable quality" is relatively objective (§2.71.1–72.3)◑. Schultz-Bergin's "grade anarchy" in philosophy saw grade inflation at self-assessment — most students assigned themselves A-range grades, inconsistent with midterm self-assessments at C+/B- (§2.103.3–103.4)●.

Sorensen-Unruh (2024), writing from within the ungrading community, argues that the movement's standard theoretical justification — self-regulated learning theory — is inadequate for its emancipatory aims because of SRL's "deficit frame." This is a significant internal critique: even the movement's own leading STEM practitioner finds its theoretical foundations insufficient (Abstract-verified, W4400447946).

The practical limitation of ungrading is structural: it requires institutional contexts where external accountability demands are minimal. Only tenured faculty at well-resourced institutions have the professional security to take pedagogical risks of this kind — a fact the volume acknowledges through Warner's testimony, a contingent faculty member who ultimately left higher education partly because the contradiction between his ungrading values and his contingent status became

irreconcilable (§2.118.5)[1]. The K-12 coverage is thin — only a handful of contributors (Kirr, Chu, Sackstein) address K-12 contexts, and the institutional constraints (mandatory state standards, external testing, parent constituencies) are substantially different and largely unengaged.

5.4 COMPETENCY-BASED ASSESSMENT

Competency-based assessment decouples evaluation from time: students advance when they demonstrate mastery of defined competencies, not when they have completed specified seat hours. Medical education has the most extensive implementation experience, following Harden's (1999) foundational articulation of outcome-based education.

Ten Cate's (2013) Entrustable Professional Activities (EPAs) — with 963 citations, the most-cited medical competency-based assessment construct — provide the structural mechanism: clinical supervisors make a trust decision about whether a trainee can perform a specific professional activity without direct oversight. The trust decision integrates multiple observations, assessment instruments, and professional judgments over time — a form of assessment that is inherently subjective, longitudinal, and expert-dependent (Abstract- verified, W2018012367).

Harden's (2016) OSCE retrospective documents 40 years of Objective Structured Clinical Examinations. The OSCE is a structured performance-based assessment that standardizes the clinical encounter through stations with trained standardized patients. Its strengths are reliability and standardization; its limitations are artificiality (the standardized patient is not a real patient) and the assessment of isolated skills rather than integrated clinical judgment. Prentice et al. (2020) document workplace-based assessment (WBA) failure modes: supervisors give pro forma satisfactory ratings (the "tick-box" problem), assessment moments are not well-timed relative to learning needs, and the burden of documentation is resisted by clinicians who see it as administrative overhead (Abstract- verified).

Ten Cate and Regehr (2018) argue against the objectivity ideal in assessment — the argument that subjective expert judgment carries information that reliability-maximizing standardization suppresses. When experts disagree about a trainee's readiness, that disagreement itself carries developmental information: it reveals areas of genuine clinical ambiguity rather than merely reflecting measurement noise. Their structural solution is programmatic assessment: aggregating many subjective expert judgments over time across contexts and assessors, with decisions made by expert committees rather than by any single assessment event (Abstract-verified, FWCI 17.92).

The medical-education experience teaches three lessons for assessment system design. First, competency-based assessment is institutionally demanding: clearly defined competencies, valid assessments, trained assessors, and flexible structures that allow variable-rate progression. Second, assessment reliability and assessment validity are in genuine tension: the most reliable assessments (standardized, structured, controlled) are often the least valid for assessing complex professional judgment. Third, implementation at scale is possible — medical education does it — but requires teacher-to-student ratios and institutional investment that most K-12 and undergraduate contexts cannot match.

5.5 PORTFOLIO ASSESSMENT

Portfolio assessment — collecting student work over time and evaluating it holistically — addresses several limitations of conventional testing. It captures growth and development, can include diverse evidence types, and can involve students in selecting and evaluating their own work.

But portfolio assessment at scale has repeatedly failed. The Vermont portfolio assessment experience — documented by Koretz et al. (1994), though this source could not be verified in OpenAlex and rests partially on training knowledge — found substantial interrater reliability problems: trained raters disagreed substantially on the same portfolio. Vermont subsequently scaled back the program. Kentucky had a similar experience. The scoring-burden problem is structural: portfolio assessment at scale requires either large quantities of trained rater time (expensive and introducing reliability concerns) or reduced scoring depth (defeating the purpose).

The honest claim: portfolio assessment is viable at small scale with trained assessors and robust moderation processes. It has repeatedly failed as a large-scale accountability mechanism. Its formative function — where reliability demands are lower — is more defensible than its summative function (Training- derived, consistent with the broader assessment literature).

6 THE ASSESSMENT-MOTIVATION TENSION

6.1 THE DEFINING UNSOLVED PROBLEM

This section addresses v1 Gap 1 — the defining unsolved problem of the assessment domain. The tension is simple to state and difficult to resolve: assessment practices that maximize learning (retrieval practice, formative feedback) operate in the same institutional space as assessment practices that damage motivation (grades, rankings, high-stakes testing). The question is whether a single assessment system can serve both functions without one undermining the other.

6.2 THE MOTIVATION EVIDENCE BASE

The evidence that grades undermine intrinsic motivation is well-established, though not without boundary conditions. Deci, Koestner, and Ryan's (1999) meta-analysis of 128 studies found that tangible, expected, performance-contingent rewards — exactly the structure of grades — reliably undermine intrinsic motivation. Verbal, informational feedback that preserves self-determination can enhance it (Training-derived, confirmed across multiple sources). Bureau et al. (2021), in a meta-analysis of 144 studies with N > 79,000, confirmed that teacher autonomy support predicts need satisfaction and self-determined motivation more strongly than any other contextual factor, and that competence is the strongest predictor of autonomous motivation, followed by autonomy (Abstract-verified, W3201580913, FWCI 33.69).

Kohn's (1993) *Punished by Rewards* provides the most extensive engagement with the anti-grades argument. His central claim is that external reward systems — including grades — do not merely fail to sustain motivation but actively undermine it, degrading both performance quality and intrinsic engagement. The mechanism is the control-versus-information distinction: rewards experienced as controlling degrade intrinsic interest regardless of their form. The effect extends beyond tangible rewards: evaluation, surveillance, deadlines, and competition all produce the same erosion (§1.12.12–1.12.14)•.

On grades specifically, Kohn examines three standard justifications — motivation, sorting, and feedback — and rejects all three. A letter grade communicates nothing actionable; "a substantive comment that does offer such information, meanwhile, gains nothing from the addition of the B+" (§1.25.6)•. Students who focus on grades shift into a performance orientation — concerned with how they compare to others — that predicts "worse memory, less creative thinking, less conceptual learning, reduced intrinsic motivation, and increased fear of failure" (§1.19.15–19.17)•. This applies to high achievers as well: "it is not only those punished by F's but also those rewarded by A's who bear the cost of grades" (Training-derived, consistent with the source).

The Eisenberger-Cameron counter-position must be engaged honestly. Cameron (2001), responding directly to Deci, Koestner, and Ryan's meta-analysis, argued that the undermining effect of tangible rewards is a "limited phenomenon" applying only to specific conditions (Abstract-verified, W2163770743). Cameron and Pierce's (1994) earlier meta-analysis had argued the same. The scholarly exchange established that the undermining effect is real for tangible, expected, performance-contingent rewards — exactly what grades are — but not universal. Unexpected rewards, task-noncontingent rewards, and informational verbal feedback do not reliably undermine

and may enhance motivation. The Deci et al. (1999) position is better supported by the subsequent literature than the Eisenberger-Cameron position, but the distinction between controlling and informational is the operational hinge, not the simple presence or absence of rewards.

6.3 WILIAM'S CONTESTED REVERSAL

Wiliam (2011) complicates the standard SDT account with a significant and contested claim: "the evidence, such as it is, suggests that causation is running in the opposite direction. Motivation is not a cause but a consequence of achievement" (§31.0.12)•, citing Garon-Carrier et al. (2016). If motivation follows achievement rather than causing it, then engineering small wins through effective assessment design is more important than protecting motivation through assessment-system reform.

This claim is in direct tension with the SDT literature, which documents causal pathways from controlling assessment to reduced intrinsic motivation. Both positions have empirical support: the SDT evidence for the undermining effect of controlling assessment is robust, and the reciprocal-effects literature showing that achievement gains produce motivational gains is also supported. The honest synthesis is that causation runs in both directions — achievement affects motivation and motivation affects achievement — and that assessment design must attend to both pathways. Wiliam is right that producing small wins matters. Kohn and the SDT theorists are right that the form in which those wins are communicated — informational versus controlling — matters for whether the motivational benefit is sustained.

6.4 THE W2-008 ENVIRONMENTAL REFRAME

W2-008's review of the curriculum-philosophy literature introduced a finding that changes the terms of the assessment- motivation debate. Watts, Duncan, and Quan (2018) shrank the marshmallow-test effect by approximately half once family background was controlled. The implication, as W2-008 develops it, is that self-regulation is not primarily a trainable individual capacity but an environmentally conditioned one. The environmental conditions that support self-regulation include the relational and motivational climate of the learning environment — which assessment design shapes directly.

This reframe has a specific implication for the assessment- motivation tension: the damage a poorly designed assessment regime does to motivation and self-regulation may be more durable and harder to counteract than the direct-training literature implies. If self-regulation develops through environmental support rather than through explicit instruction, then an assessment environment that is chronically controlling — that communicates through grades, rankings, and sanctions rather than through informational feedback — does not merely suppress motivation in the moment. It may degrade the environmental conditions under which self-regulation develops, with consequences that persist beyond the immediate assessment event.

No longitudinal study has tested this specific prediction directly. The longitudinal evidence on assessment effects is, as v1 Gap 8 noted, essentially nonexistent — multi-year studies comparing different assessment regimes on motivation and self-regulation trajectories do not exist. But the theoretical convergence between the SDT undermining effect (controlling assessment reduces intrinsic motivation), the environmental self-regulation finding (self-regulation is supported by environmental conditions, not trained in isolation), and the clinical evidence from East Asian examination cultures (see Section X) is strong enough to take seriously: assessment regime design may be a character- formation mechanism, for good or ill, whether or not it intends to be.

6.5 WHAT CAN AND CANNOT BE RESOLVED

The assessment-motivation tension cannot be fully resolved given current evidence. What can be said, with convergent support from the cognitive-psychology literature, the motivation literature, and the practitioner traditions, is this:

Six convergent design principles emerge from the assessment-motivation scorecard (this agent's Session 2 sharpening artifact, drawing on all reading notes):

1. *Do not grade formative work simultaneously with providing feedback.* The Butler (1988) finding — confirmed by Wiliam's (2011) analysis, by Kluger and DeNisi's (1996) FIT mechanism, and by the SDT undermining-effect evidence — is that grade and feedback fight for attentional resources, and the grade wins.

2. *Use criterion-referenced rather than norm-referenced grading.* Norm-referencing maximizes social comparison, the most ego-threatening and least informative reference frame. The cognitive and motivation literatures agree with unusual clarity that norm-referenced grading is the worst available option on both dimensions (Marzano, 2000; Kohn, 1993; Kluger & DeNisi, 1996).

3. *Keep formative assessment low-stakes and unrecorded.* Retrieval practice without grading stakes is the most robustly supported formative mechanism (Wiliam, 2011; Yang et al., 2021).

4. *Make self- and peer assessment central, not supplementary.* Both produce learning and autonomy benefits that teacher- only feedback cannot replicate (Carless & Boud, 2018; Nicol, 2020)[○].

5. *Use rubrics as orientation devices before work begins, not as scoring instruments after.* Sadler (2008, 2010) and Carless and Chan (2016) converge on this: rubrics work better for feed-up than for evaluation.

6. *Prefer delayed process-level feedback over immediate task-level correction where the domain allows.* Conceptual development benefits from productive struggle before feedback (Wiliam, 2011, citing Bjork and Mullet et al., 2014).

What remains unresolved: Whether any assessment system can fully serve both the informational function (supporting learning) and the institutional accountability function (certifying competence for external stakeholders) without one compromising the other. The ungrading movement argues the two are incompatible and eliminates the accountability function. The SBG movement argues they can be reconciled through criterion-referencing and disaggregation. The medical-education experience suggests that programmatic assessment — aggregating many low-stakes observations into high-stakes decisions made by expert committees — is the closest available operational model. No controlled comparison of these approaches on both learning and motivational outcomes, at scale, over multiple years, exists.

7

FEEDBACK IN ILL-STRUCTURED DOMAINS

7.1 WHY THIS CHAPTER MATTERS

v1 identified feedback for ill-structured tasks as Gap 2 — a practical blind spot in the assessment literature. The mainstream feedback frameworks (Hattie & Timperley, 2007; Kluger & DeNisi, 1996) were developed primarily from studies of well-structured tasks where there is a clear standard of correctness. Applied Pedagogy's curriculum will necessarily include ill-structured tasks — writing, ethical reasoning, design thinking, clinical judgment — and the standard feedback models provide limited guidance for these domains.

This section engages four domain-specific feedback traditions that have developed substantial practical knowledge about how to give feedback on ill-structured work: writing-center pedagogy, studio-critique in design education, script- concordance testing in clinical reasoning, and the evaluative- judgment tradition in artistic and academic performance. It maps each against the Hattie-Timperley and Kluger-DeNisi frameworks to identify what transfers and what does not, and it draws on Sadler's (2010) guild-knowledge concept and Carless and Boud's (2018) feedback-literacy framework as bridging theories.

7.2 WRITING

The writing-center tradition, as described by North (1987)[○], represents the longest-running practice of feedback on ill-structured work in education. Its distinctive features are: one-on-one conferencing rather than written comments on finished work; process orientation (responding to writing in progress, not to a completed product); non-evaluative stance (writing-center tutors do not grade — they respond, generating dialogue); a "higher-order concerns first" hierarchy (meaning, argument, and structure before grammar and mechanics); and writer ownership (the tutor does not fix the writing; the writer decides what to change).

The writing-center tradition's higher-order-concerns-first hierarchy converges with Hattie and Timperley's process-level prioritization: both agree that feedback on meaning and argument should precede feedback on surface features. The non-evaluative stance aligns with Kluger and DeNisi's prediction that feedback is most effective when it keeps attention at the task-learning level — by removing grades from the feedback encounter, writing centers structurally prevent the attention shift that FIT identifies as the primary mechanism of feedback harm.

Where the traditions diverge is more revealing than where they converge. The Hattie-Timperley model treats feedback as a three-question sequence organized around closing a gap to a known goal. In writing, however, the goal itself is frequently emergent — the writer discovers what they want to say during the writing process, and the tutor's job is partly to help the writer discover their own goal rather than close the gap to a pre-specified one. Sadler's (1989) gap-closing model assumes the goal is known; the writing tradition recognizes that in writing, the goal evolves. This is a genuine structural difference between well-structured and ill-structured domain feedback that the standard frameworks do not address.

Additionally, the writing-center tradition's emphasis on dialogue — feedback as co-construction between writer and reader — has no equivalent in the Hattie-Timperley model, which retains

an implicit transmission structure. Carless and Boud's (2018) dialogic peer-feedback mechanism comes closer, but even their framework is primarily concerned with how students use feedback rather than with feedback as joint inquiry into an evolving text.

Danielewicz and Elbow (2009), in what the writing-studies field most frequently cites on grading contracts (155 citations, W2018012367)○, provide an alternative assessment structure: the contract specifies requirements for a given grade in terms of process (number of drafts, participation in peer review, minimum-length criteria) rather than quality. Quality feedback is separated entirely from grade determination — the contract handles the institutional accountability function while leaving the feedback function free from evaluative contamination.

7.3 DESIGN EDUCATION

The design studio critique is structurally unlike any feedback instrument in the cognitive-psychology literature. Goldschmidt, Hochman, and Dafni (2010), analyzing three architecture studio crits through protocol analysis and linkography, found that the most productive crit exchanges are generative rather than evaluative: the teacher proposes a new design direction, the student responds, the teacher elaborates — feedback as the occasion for generating new design possibilities rather than evaluating existing ones (Abstract-verified, FWCI 34.96).

The design critique challenges the Hattie-Timperley level taxonomy in a fundamental way. The task/process/self distinction breaks down in the crit: a comment about structural engineering is simultaneously a comment about the design process and about the designer's professional judgment. The levels presuppose separability; the design crit shows they typically cannot be separated. Dannels, Housley Gaffney, and Norris Martin (2008) found that critique feedback implicitly assesses five communication competencies — interaction management, demonstration of design evolution, transparent advocacy of intent, explanation of visuals, and staging of the performance — that constitute what it means to reason as a designer (Abstract-verified, FWCI 34.96).

Most importantly, design problems have no correct answers. "Good" design is contested among experts. Kluger and DeNisi and Hattie and Timperley both assume feedback is against a known standard the teacher holds. The design crit's standard is the contested judgment of a design community — not a pre-specified criterion. This is the most fundamental structural challenge the ill-structured-domain traditions pose to mainstream feedback frameworks.

7.4 CLINICAL REASONING

Clinical education has produced the most formalized instruments for feedback on ill-structured reasoning. Script Concordance Testing (SCT), documented by Fournier, Demeester, and Charlin (2008) and evaluated by Lubarsky et al. (2011), assesses clinical reasoning under uncertainty by presenting a clinical scenario where new information could change the probability of a diagnosis. The scoring key is not a preset answer but the statistical distribution of responses from an expert clinician panel — meaning the standard itself is socially calibrated consensus, not a deterministic answer (Abstract-verified, FWCI 6.57 and 8.86).

Lubarsky, Dory, Audétat, Custers, and Charlin (2015) connect SCT to illness-script theory: clinical expertise consists in well-calibrated knowledge networks bundling symptoms, mechanisms, and management plans, and feedback that accelerates expertise must engage the structure of the learner's scripts, not just the accuracy of their conclusions (Abstract-verified, FWCI 4.79). This connects directly to W2-009's treatment of the skill-to-judgment transition at Layer 3 of the

competence stack: feedback for judgment looks different from feedback for skill because judgment involves probabilistic weighting under uncertainty, not deterministic correctness.

Ten Cate and Regehr's (2018) argument for the value of subjective expert judgment against the objectivity ideal applies here with particular force. The reliability-validity inversion they identify — excessive pursuit of inter-rater reliability distorts assessment by suppressing legitimate expert judgment variance — is entirely absent from the feedback literature, which treats inter-rater agreement as an unqualified measurement virtue (Abstract-verified, FWCI 17.92). In clinical reasoning, expert disagreement carries developmental information: it reveals genuine clinical ambiguity. Programmatic assessment — aggregating many subjective judgments over time — is their structural solution.

7.5 THE EVALUATIVE-JUDGMENT TRADITION

Eisner's (1976) connoisseurship-and-criticism model provides the most theoretically developed account of feedback in artistic and complex academic performance contexts (Abstract- verified, FWCI 19.27). Connoisseurship is the perceptual capacity to discriminate among levels of quality, developed through sustained domain immersion. Criticism is the art of making connoisseurship public — translating private perception into communicable language. Feedback on complex work is educational criticism in this sense: making expert evaluative capacity visible and learnable.

Sadler (2010) names this "guild knowledge" — the accumulated practical and evaluative wisdom that expert practitioners hold, often implicitly, about what constitutes excellent work. Students develop it through sustained exposure and guided comparison, not through rule-learning (Abstract-verified). Carless and Chan (2016) document the "withholding strategy" in exemplar-based dialogic feedback: when a teacher withholds evaluation and invites student reasoning about an exemplar, students must generate their own quality judgments rather than copying the teacher's. The pedagogical value depends on how exemplars are discussed, not merely on exposure (Abstract- verified, FWCI 46.50). To, Panadero, and Carless (2021) add the draft-first principle: students who attempt the task before seeing exemplars produce stronger evaluative development than those who see exemplars before making their own attempt (Abstract-verified, FWCI 12.27).

7.6 CROSS-WALK: WHAT TRANSFERS AND WHAT DOES NOT

Three claims converge across all four traditions and are absent from or underspecified in the mainstream feedback literature:

First, expert judgment is irreducibly subjective and legitimate. Ten Cate and Regehr argue this for clinical assessment; Goldschmidt et al. assume it for design; Sadler argues it against rubric orthodoxy; Eisner grounds it in aesthetic theory. The mainstream feedback frameworks — built on the assumption of a knowable standard — cannot accommodate this convergent finding.

Second, quality standards are tacit and require apprenticeship to transmit. Sadler's guild knowledge, Eisner's connoisseurship, and the illness-script framework all name the same structural fact: expert evaluative judgment develops through sustained exposure, practice, and guided comparison in communities of practice. It is not reducible to explicit criteria.

Third, feedback must engage the learner's reasoning process, not just the output. The SCT mechanism engages clinical reasoning directly. The design crit's generative function creates new design moves rather than evaluating artifacts. Writing-center conferencing addresses the writing process, not the finished text.

What does not transfer: the cognitive-psychology assumption that there exists a knowable, agreed-upon standard against which performance can be measured. This assumption holds for well-structured tasks and fails — predictably and structurally — for ill-structured ones. For curriculum design, this means that ill-structured-domain assessment requires a different apparatus from well-structured-domain assessment: community calibration (SCT-style expert panels, studio-critique traditions, peer review with trained evaluators), exemplar-based orientation (Carless & Chan, 2016; To, Panadero & Carless, 2021), and holistic expert judgment rather than analytic rubric scoring.

8

FEEDBACK FOR JUDGMENT

8.1 THE LAYER 3 PROBLEM

W2-009's review of competence formation identified a transition point between Layer 2 (skill) and Layer 3 (judgment) that has specific implications for feedback design. Skill feedback addresses whether the performance met defined criteria — correctness, completeness, efficiency. Judgment feedback addresses whether the performer weighed competing considerations appropriately under uncertainty — a fundamentally different construct. Feedback for skill says "you did this wrong; here's how to do it right." Feedback for judgment says "your weighting of these competing factors was uncalibrated; here's how an expert thinks about the tradeoff."

The cognitive-psychology feedback literature — Hattie and Timperley's four levels, Kluger and DeNisi's FIT — was not designed to address judgment under uncertainty. Task-level feedback assumes a correct answer exists. Process-level feedback assumes transferable strategies can be identified. Self-regulation feedback assumes the learner can monitor their approach against a known standard. For judgment, none of these assumptions holds cleanly: the "correct" answer is probabilistic, the strategies are contextual rather than transferable, and the standard is the contested assessment of expert communities rather than a fixed criterion.

8.2 NATURALISTIC DECISION MAKING

The naturalistic decision-making literature (Klein, 1998; the recognition-primed decision model) provides the theoretical foundation for understanding how experts make judgments. Klein's model proposes that experienced decision- makers do not evaluate options by comparing alternatives against criteria — they recognize situations through pattern matching against previous experience and mentally simulate their first plausible response. If the simulation works, they act; if not, they modify or consider the next most plausible option (Training-derived, widely cited in the expertise literature).

The implication for feedback is that developing judgment requires sustained exposure to authentic decision situations — with feedback that calibrates the learner's pattern-recognition and simulation capacity. This is not the kind of feedback the standard models describe. It is closer to what the clinical-reasoning tradition practices through case presentation and what the design tradition practices through desk crits — feedback embedded in authentic judgment tasks, delivered by a more experienced practitioner who makes their own reasoning visible.

W2-009's treatment of this transition proposes that the development of evaluative capacity — the ability to distinguish good judgment from poor judgment — is the core mechanism. Eisner's connoisseurship applies here as well: the expert can perceive quality distinctions that the novice cannot, and the feedback task is to make those perceptions accessible. This review cites W2-009's Layer 3 analysis rather than duplicating it and focuses on the assessment-specific implication: if judgment develops through apprenticeship and pattern exposure rather than through explicit instruction and corrective feedback, then the assessment instruments appropriate for Layers 1–2 (quizzes, rubrics, demonstrations) are structurally insufficient for Layer 3. Assessment of

judgment requires something closer to the clinical tradition's programmatic assessment — multiple observations over time, expert calibration, and consensus-based trust decisions.

8.3 SCRIPT-CONCORDANCE TESTING AS FORMALIZED JUDGMENT FEEDBACK

SCT, discussed in Section VII, is the most formalized instrument for providing feedback on judgment under uncertainty. Its distinctive feature — scoring against the distribution of expert responses rather than against a preset correct answer — makes it a genuine judgment assessment rather than a knowledge assessment dressed up in clinical language. The expert panel's distribution of responses is the standard, and a trainee who deviates from it is not necessarily wrong — they are uncalibrated. The feedback is: "Your probability estimate diverges from expert consensus at this point." The learning value comes from what the trainee does with that information: examining their own reasoning, identifying why their weighting differed, and recalibrating their illness scripts accordingly.

SCT suggests a general principle for judgment feedback: calibration against expert communities, not correction against fixed answers. This principle is transferable beyond clinical reasoning to any domain where judgment under uncertainty is the target competence — including ethical reasoning, policy analysis, entrepreneurial decision-making, and artistic evaluation.

9

AI-ASSISTED FEEDBACK AND ASSESSMENT

9.1 THE STATE OF THE EVIDENCE

The AI-assisted feedback literature is approximately 24 months old in the GPT-4 era. This section engages it at its current state of development — honestly about what has been shown, what has not been tested, and where the emerging findings are most concerning.

The evidence can be organized using the Hattie-Timperley four-level framework, and the pattern that emerges is clear: AI feedback is competitive at the task level, partially competitive at the process level, consistently fails or causes harm at the self-regulation level, and is structurally absent from the self level.

9.2 TASK LEVEL: AI IS COMPETITIVE

AI feedback at the task level — correcting surface errors, identifying structural completeness, marking grammar and citation format — is the most technically mature and empirically supported function. Meyer et al. (2023), in an RCT of 459 German upper-secondary EFL students, found that GPT-3.5-turbo feedback on argumentative essays produced small but positive gains in revision quality (d = 0.19) relative to no feedback, with moderate positive motivational effects (d = 0.34–0.36) (Abstract-verified). Escalante et al. (2023), in a quasi-experimental study of Hawaiian ENL university students over six weeks, found no significant difference in writing-quality outcomes between GPT-4 feedback and human-tutor feedback — a null result consistent with AI competitiveness for surface-level correction (Abstract- verified).

The scalability advantage is genuine. For large classes where the realistic alternative is no feedback on formative drafts, AI task-level feedback is clearly better than nothing. This is a meaningful finding for the teacher-bandwidth constraint that motivates AI-feedback research.

9.3 PROCESS LEVEL: PARTIALLY COMPETITIVE

At the process level — feedback about the strategies used to perform a task — AI's competitiveness weakens. Banihashem et al. (2024), comparing ChatGPT and peer feedback on academic essays, found that ChatGPT feedback is characteristically descriptive (task-level) while peer feedback is characteristically diagnostic (process-level). Peers outperform ChatGPT on the diagnostic dimension — identifying what is wrong with the reasoning, not just what is wrong with the text (Abstract-verified, W4394710139). Escalante et al. (2023) confirm that students prefer human feedback for "contextually aware" responses: understanding why an argumentative choice was made and what alternative strategy would work better is process-level feedback requiring interpretive capacity that current LLMs do not reliably deploy.

9.4 SELF-REGULATION LEVEL: AI CONSISTENTLY FAILS

Fan et al. (2024) provide the most consequential finding in the AI-feedback literature. In an RCT with 117 participants, ChatGPT users showed performance gains on the essay task but *no gains in*

knowledge or transfer — and significantly reduced engagement with self-regulation sub-processes (evaluation, monitoring, orientation) compared to learners supported by a human expert, writing analytics tools, or no support at all. The AI performs the external monitoring function, crowding out internal monitoring. The performance-learning dissociation is the core finding: AI does what self-regulation is supposed to do, externally, so the learner does not develop the internal capacity (Abstract-verified, W4405211386, FWCI 118.3).

Darvishi et al. (2023) confirm the pattern in a larger-scale study (RCT, N = 1,625, 10 courses). AI prompts during peer review improved the quality of feedback given while prompts were present; when removed, students reverted substantially — demonstrating that AI did not build self-regulation capacity but replaced it. The hybrid condition (AI prompts plus self-monitoring checklist simultaneously) was not more effective than AI alone — showing that simply layering a self-monitoring requirement on top of AI feedback does not prevent the dependency that displaces self-regulation (Abstract-verified, W4389210056, FWCI 91.1).

Bauer et al. (2025) provide the analytical vocabulary for this pattern through their ISAR taxonomy: Inversion (AI harms learning by reducing cognitive engagement), Substitution (AI replaces instruction with equivalent effectiveness), Augmentation (AI adds support beyond what instruction alone provides), and Redefinition (AI enables qualitatively new tasks). Fan et al. and Darvishi et al. document inversion at the self-regulation level. No confirmed example of redefinition exists in the feedback literature yet (Abstract-verified, W4409772554, FWCI 77.5).

9.5 THE METACOGNITIVE-LAZINESS MECHANISM

Fan et al.'s (2024) metacognitive laziness operates through a straightforward mechanism. When AI feedback provides fully formed, actionable corrections, the student can comply without understanding why the correction was needed. The metacognitive processing that would develop evaluative capacity — the internal monitoring and judgment that constitute self- regulation — is bypassed by the immediacy and specificity of AI feedback. This begins at the task level (where compliance without understanding is easiest), intensifies at the process level (where AI diagnosis relieves the learner of diagnostic effort), and culminates at the self-regulation level (where AI monitoring replaces internal monitoring).

The concern is strongest under three jointly identified conditions: (1) complex, ill-structured tasks that require precisely the evaluative and monitoring processes AI displaces; (2) immediately actionable AI feedback that can be applied without comprehension; and (3) novice learners without external criteria against which to evaluate AI suggestions (Kasneci et al., 2023, on AI literacy; Bauer et al.'s moderating- factors argument). The concern is weakest when AI feedback is used as a first-pass that the learner is then asked to evaluate critically — a design that preserves metacognitive processing.

9.6 WHAT AI FEEDBACK DOES AND DOES NOT SOLVE

The honest synthesis: AI feedback is better than no feedback (Meyer et al., 2023). It may be equivalent to average human feedback for surface outcomes (Escalante et al., 2023). It is reliably worse than good human feedback for building self-regulation capacity (Fan et al., 2024; Darvishi et al., 2023). The literature is 24 months old and evolving rapidly; these findings are provisional.

The most defensible design principles as of April 2026 are: (1) use AI for task-level first-pass feedback on written work, freeing teacher time for process- and self-regulation-level feedback; (2) layer explicit metacognitive-engagement prompts over AI feedback, requiring students to evaluate

each suggestion before applying it (not yet empirically validated but theoretically predicted to mitigate the Fan et al. effect); (3) use AI to scaffold peer feedback production rather than to replace peer feedback, as in the Guo et al. (2024) design direction; (4) plan fading trajectories — progressively withdrawing AI scaffolding while monitoring whether self-regulation fills the gap; and (5) restrict AI feedback to task and limited process levels for novice learners on complex ill-structured tasks, where the metacognitive- laziness risk is highest.

10 CROSS-CULTURAL VARIATION IN ASSESSMENT

10.1 WHY CULTURAL CONTEXT MATTERS

v1 Gap 6 noted that the assessment literature is overwhelmingly drawn from Western, English-language contexts. This section engages the non-English assessment research traditions that have developed distinct perspectives on formative assessment, examination culture, and the assessment-motivation tension.

10.2 EAST ASIAN EXAMINATION CULTURES

The Chinese gaokao, Korean suneung, and Japanese hensachi systems represent the limit cases of high-stakes assessment — natural stress tests of the SDT prediction that controlling assessment damages motivation.

Kirkpatrick and Zang (2011), studying Chinese high school students in Yunnan Province, documented the motivational and psychological damage of exam-oriented pedagogy. Students reported viewing education as "nothing more than merely passing examinations" — a framing that evacuates the intrinsic value of learning entirely. Creativity, sense of self, and psychological health were all impaired (Abstract-verified, W2140830768). The sample was small (n = 43) and from a single school, but the findings are consistent with the broader backwash literature and with the larger Liu and Helwig (2020) study.

Liu and Helwig (2020), through in-depth clinical interviews with Chinese secondary school graduates, found that students from all backgrounds — urban elite and rural disadvantaged — reported experiencing psychological conflict between their understanding of education's intrinsic value and the extrinsic pressure of the gaokao. This universal motivational conflict finding complicates the cultural-specificity objection to SDT: even in cultures with strong collectivist examination traditions, the damage of purely extrinsic, high-stakes assessment is felt (Abstract-verified, W3017868667, FWCI 5.15).

Chung and Park (2024), studying Korean high school students, documented that the harms of the suneung ecosystem extend beyond motivation to clinical-range mental health: anxiety disorders, depression, and in extreme cases suicidality. South Korea has the highest rate of student suicide among OECD nations. The suneung is norm-referenced (percentile scores, not raw scores), meaning every student's outcome directly depends on every other student's — a structural feature that maximizes competitive stress. This is not merely a motivational cost in the SDT sense; it is a clinical harm that changes the assessment design question from "does this undermine intrinsic motivation?" to "does this produce psychological damage?" (Abstract-verified, W4391260222, FWCI 5.58).

10.3 THE GERMAN REFERENCING FRAMEWORK

Rheinberg (2001) developed the Bezugsnormen (reference-norm) framework, which distinguishes three assessment reference frames with different motivational and cognitive consequences: social reference norms (comparing the student to peers), criterion reference norms (comparing against

absolute standards), and individual reference norms (comparing against the student's own prior performance). Rheinberg's research shows that teachers who emphasize individual reference norms — ipsative assessment — produce better motivational outcomes: students develop stronger effort attributions, higher self-concepts of ability, and reduced test anxiety (Abstract-verified).

Klieme et al. (2003) describe the German post-PISA Bildungsstandards reform, which introduced competence-based assessment at the national level — a systematic alternative to the traditional Leistungsbeurteilung (achievement grading) system. The reform represents one of the most ambitious attempts to shift from norm-referenced to criterion-referenced assessment at national scale (Abstract-verified).

10.4 THE FRENCH ÉVALUATION FORMATIVE TRADITION

The French-language formative assessment tradition — particularly the work of Morrissette (2010/2014) and Mottier Lopez and Laveault (2008/2014) — developed independently from and partly in parallel with the Anglophone Black and Wiliam tradition. Morrissette's systematic review, published on the open-access Érudit platform, provides a French-language synthesis of formative assessment theory and practice. Mottier Lopez and Laveault introduce the distinction between "régulation interactive" (real-time adjustment during instruction) and "régulation rétroactive" (post-hoc adjustment based on assessment results) — a finer-grained temporal distinction than the Anglophone literature typically makes (Abstract-verified).

10.5 THE VYGOTSKYAN DYNAMIC ASSESSMENT TRADITION

Poehner and Lantolf (2005), working from the Vygotskyan tradition, propose Dynamic Assessment (DA) as an alternative to conventional formative assessment. DA deliberately intervenes during the assessment task: the assessor provides graduated mediation — hints, prompts, explicit instruction — and the learner's responsiveness to mediation becomes the central datum, not their independent performance level. A learner at level 5 who, with a minimal prompt, jumps to level 9 is developmentally different from a learner at level 7 who plateaus despite assistance. Conventional assessment cannot detect this; DA is designed precisely to surface it (Abstract-verified, W2096750265, FWCI 8.57).

Poehner and Lantolf argue that Black and Wiliam's formative assessment framework could be strengthened by DA principles: the key missing element is a theorized account of *how* teachers should intervene, not just *that* they should. DA's graduated mediation hierarchy provides this missing specification. The underlying premise — that assessment should reveal developmental potential, not current attainment — is distinctly different from the Anglo-American measurement tradition and represents a genuine alternative theoretical foundation for formative assessment.

10.6 THE FINNISH NO-TESTING REGIME

Finland famously defers standardized testing until age 16. The Finnish assessment literature — both Finnish-language and English-language — documents formative assessment practices embedded in teacher autonomy and professional trust rather than in external accountability systems. Hendrickson (2012) and others describe an assessment culture that relies on teacher-generated classroom assessment, without national examinations driving backwash effects, until the matriculation examination at the end of upper secondary school. The contrast with East Asian examination cultures is stark: Finland's assessment environment is among the least controlling in

OECD nations, and its educational outcomes — while recently declining from their PISA peak — remain strong (Training- derived, consistent with the comparative education literature).

10.7 WHAT CROSS-CULTURAL EVIDENCE ADDS

The cross-cultural evidence adds three things to the assessment-motivation synthesis:

First, the assessment-motivation tension is not culturally specific. Liu and Helwig's finding that Chinese students experience the intrinsic-extrinsic conflict despite strong collectivist examination traditions suggests that the SDT account has broader applicability than the Western-specificity objection implies.

Second, the harm scales sharply with stakes, singularity, and norm-referencing. The East Asian evidence shows that the damage is not linear — it concentrates at the extreme of high-stakes, norm-referenced, singular examinations and escalates from motivational damage to clinical psychological harm.

Third, alternative assessment cultures exist and function. The Finnish model demonstrates that high-quality educational outcomes are possible without the accountability-testing apparatus that drives assessment design in most of the English-speaking world. This does not prove that removing testing improves outcomes (Finland's educational culture involves many other distinctive features), but it refutes the claim that external high-stakes testing is necessary for quality.

11

PRACTICAL IMPLICATIONS FOR CURRICULUM DESIGN

11.1 FROM EVIDENCE TO SYSTEM

The evidence reviewed in Sections II–X converges on a set of design principles that are more specific and more grounded than v1's seven principles. This section specifies what a concrete assessment system for Applied Pedagogy's curriculum would look like — what formative mechanisms it uses, how feedback is delivered, how retrieval practice is scheduled, what grading approach is appropriate, how AI-assisted feedback is and is not used, and what the expected motivational consequences are.

11.2 ASSESSMENT ARCHITECTURE

The assessment system has three tiers, each serving a different function:

Tier 1: Daily Formative Assessment (ungraded, unrecorded). Brief retrieval practice — 3–5 minutes at the start of every learning session. Free-recall or short-answer format, spaced across topics to leverage both the testing effect and the spacing effect. Immediate corrective feedback provided. No grades, no recording, no stakes. The purpose is purely formative: strengthening memory through retrieval and generating evidence (for both teacher and student) of what has and has not been learned. The cognitive evidence (Yang et al., 2021, g = 0.49; Roediger & Karpicke, 2006) and the motivational evidence (Wiliam, 2011; SDT) converge: this is the most robustly supported formative mechanism available.

Tier 2: Weekly Formative Assessment (feedback-rich, revision- oriented, low-stakes). More substantial formative work — problem sets, short writing assignments, peer review exercises, self-assessment activities. Detailed process-level feedback is provided within 24–48 hours, focused on strategies and reasoning rather than on surface correctness. Students have a structured revision opportunity: every piece of feedback is followed by an opportunity to act on it. If institutional context requires grades, they are based on improvement or completion rather than first-attempt accuracy, and they are delivered separately from feedback (Butler, 1988; Wiliam, 2011).

Peer feedback is central, not supplementary. Students regularly evaluate each other's work against criteria, which develops both the evaluator's understanding (the guild- knowledge mechanism that Sadler, 2010, describes) and the recipient's feedback literacy (Carless & Boud, 2018). Self-assessment is structured and scaffolded: students evaluate their own work against explicit criteria, compare their self-assessment to peer and teacher assessments, and gradually develop calibration accuracy. Panadero, Jönsson, and Botella (2017) show that self-assessment improves self-regulation when moderated by training and calibration verification — but novice learners, particularly low performers, systematically overestimate their performance without scaffolding (Abstract- verified, W2745915040).

Tier 3: Periodic Summative Assessment (criterion-referenced, low-frequency, multiple pathways). Summative assessments occur at unit boundaries or natural milestones. They are criterion-referenced (against defined competency standards, not against peers) and disaggregated by standard (specific competencies rated separately rather than collapsed into a single grade). Students who demonstrate mastery proceed; students who do not receive additional instruction and another

opportunity. No single assessment event is catastrophically high-stakes — the programmatic-assessment principle from medical education (ten Cate & Regehr, 2018) applies: trust decisions are made by aggregating multiple observations over time.

11.3 FEEDBACK DESIGN BY DOMAIN TYPE

For well-structured domains (mathematics, factual knowledge, procedural skills): Task-level corrective feedback is appropriate and effective for error correction. Process- level feedback should address strategy selection and error- detection approaches. Retrieval practice is the primary formative mechanism. AI-assisted feedback can serve as a first-pass task-level filter, freeing teacher time for process-level feedback. The testing-effect evidence applies most directly here.

For ill-structured domains (writing, design, ethical reasoning, clinical judgment): Process-level and self- regulation-level feedback predominate. Rubrics serve as orientation devices before work begins, not as scoring instruments after. Exemplar-based feedback (Carless & Chan, 2016; To, Panadero & Carless, 2021) develops evaluative capacity: students examine and discuss examples of quality work before attempting their own, generating internal feedback through comparison. Peer feedback is especially valuable in ill-structured domains because it develops the evaluative judgment that transfer requires. The writing-center model — one-on-one conferencing on work in progress, focused on higher-order concerns first — is the gold standard where staffing permits.

AI-assisted feedback in ill-structured domains should be restricted to task-level surface correction (grammar, citation format, structural completeness). Process-level and self- regulation-level feedback in ill-structured domains requires human judgment — the interpretive capacity that current LLMs do not reliably deploy (Banihashem et al., 2024; Escalante et al., 2023). The metacognitive-laziness risk (Fan et al., 2024) is highest precisely in ill-structured domains where students most need to develop their own evaluative capacity.

11.4 FEEDBACK DESIGN BY LEARNER LEVEL

For novice learners: Task-level corrective feedback is appropriate and necessary. Specific, directive feedback tells learners what to fix and how. Self-assessment requires scaffolding — novices lack the evaluative capacity for accurate self-assessment (Eva & Regehr, 2005; Panadero et al., 2017). AI feedback at the task level is a reasonable supplement when teacher feedback is unavailable.

For developing learners: The balance shifts toward process-level and self-regulation feedback. Feedback becomes progressively less directive and more prompting — asking questions, highlighting areas for attention, and encouraging self-diagnosis. The goal is to develop internal evaluative capacity (Nicol's, 2020, internal-feedback framework). AI scaffolding should be faded: progressively withdrawing AI support while monitoring whether self-regulation fills the gap (Darvishi et al., 2023).

For advanced learners: Self-regulation feedback is primary. The learner should be developing the capacity to evaluate their own work against expert standards — the connoisseurship that Eisner (1976) describes. Peer feedback and calibration against expert communities (SCT-style mechanisms, studio-critique traditions) replace teacher- directed feedback as the primary mechanisms. AI feedback is restricted to the first-pass filter role; advanced learners should evaluate AI suggestions against their own judgment, not accept them uncritically.

11.5 THE GRADING QUESTION

For Applied Pedagogy's curriculum, the evidence supports a grading approach that draws on SBG principles while acknowledging the evidence limitations:

- Grades are criterion-referenced and disaggregated by competency. No single-letter aggregate grades. - Grades reflect most-recent evidence of competency, not averaged performance over time (Wiliam's power-law approach, §24.0.21)●. - Non-academic factors (effort, attendance, behavior) are separated from competency assessments. - Formative work is never graded simultaneously with feedback. - Grades are delivered separately from feedback, with a temporal buffer that allows students to engage with feedback before receiving evaluative information. - Students participate in self-assessment as a regular practice, with calibration verified against teacher and peer assessments.

This is not full ungrading — the institutional accountability function is preserved through criterion-referenced competency ratings. It is not traditional grading — the most damaging practices (norm-referencing, averaging, combining academic and non-academic factors, grading formative work) are eliminated. It is closest to Wiliam's vision of competency-grid grading plus the SBG practitioners' operational frameworks.

11.6 AI-ASSISTED FEEDBACK IN THE SYSTEM

AI feedback enters the system in three roles:

1. **Task-level first-pass filter.** AI reviews student work for surface errors (grammar, citation, structural completeness) before human feedback. This frees teacher time and cognitive resources for process- and self-regulation-level feedback.

2. **Peer-feedback scaffold.** AI supports students in generating better peer feedback — not replacing peer review but raising the quality ceiling (Guo et al., 2024 design direction).

3. **Self-assessment calibration tool.** Students compare their self-assessment against AI assessment as one data point (alongside peer and teacher assessments) for calibration.

AI does not enter the system as the primary or sole feedback source for any learning activity. The metacognitive-laziness concern (Fan et al., 2024) requires that every AI feedback encounter is paired with a metacognitive-engagement prompt: "For each AI suggestion, decide whether to apply it and explain why." This is not yet empirically validated but is the intervention theory predicts will mitigate the dependency effect.

12 CLOSING ASSESSMENT

12.1 CONFIDENCE LEVELS

High confidence — build on these: - The testing effect translates to classrooms (g ≈ 0.49, Yang et al., 2021). One of the most replicated findings in cognitive psychology. - Feedback effectiveness depends on content (task and process levels), not logistics (timing, mode). Supported by multiple meta-analyses spanning three decades (Kluger & DeNisi, 1996; Wisniewski et al., 2020). - Grades cancel the learning benefit of feedback comments (Butler, 1988; Wiliam, 2011). Not directly replicated but consistent with FIT and SDT mechanisms. - Tangible, expected, performance-contingent rewards (including grades) undermine intrinsic motivation (Deci, Koestner & Ryan, 1999). Robust meta-analytic finding with boundary conditions. - Formative assessment improves learning when well-implemented (d ≈ 0.20–0.30 in rigorous studies). Direction is robust; magnitude is modest.

Medium confidence — proceed with caution: - Standards-based grading is more informational than traditional grading. Theoretically strong; empirically thin. - Peer feedback provides learning benefits for evaluator and recipient. Promising but conditional on training and structure. - Feedback literacy is teachable. Well-articulated conceptually (Carless & Boud, 2018); intervention evidence developing. - AI feedback at the task level is competitive with human feedback for surface correction. Evidence is 24 months old. - Self-assessment improves self-regulation when scaffolded. Supported by meta-analysis (Panadero et al., 2017) with important boundary conditions (calibration accuracy, novice miscalibration). - Ill-structured-domain feedback requires community calibration and exemplar-based methods. Supported by convergent evidence across four traditions; limited experimental confirmation.

Low confidence — note but do not center: - Ungrading produces better outcomes than traditional grading. Strong motivational logic; almost no comparative evidence. - Competency-based assessment is superior to time-based progression. Theoretically compelling; evidence mostly from medical education. - Portfolio assessment can substitute for conventional testing at scale. Repeatedly failed in large-scale implementations. - AI feedback with metacognitive prompts mitigates the laziness effect. Theoretically predicted; not yet tested. - The assessment-motivation tension is more severe when self- regulation is understood as environmentally conditioned. An important theoretical synthesis (W2-008) without direct empirical test.

12.2 WHAT V2 RESOLVED THAT V1 COULD NOT

v1 named the assessment-motivation tension as the defining unsolved problem. v2 has substantively narrowed it — not resolved it, but specified what the evidence supports and where it remains genuinely thin. The six convergent design principles in Section VI, the domain-specific feedback guidance in Section VII, and the concrete assessment architecture in Section XI represent specific, implementable prescriptions that v1 could not provide because it had not engaged the primary texts at depth.

v1 identified feedback for ill-structured tasks as a gap. v2 has filled that gap with substantive engagement across four domain-specific traditions (writing, design, clinical reasoning, artistic performance), identified what transfers from the mainstream feedback frameworks and what does

not, and specified what ill-structured-domain assessment requires that well-structured-domain assessment does not.

v1 flagged AI-assisted feedback as an area for future investigation. v2 has mapped the evidence against the Hattie-Timperley levels, identified the metacognitive-laziness mechanism as the central concern, and specified defensible design principles for AI feedback integration.

12.3 WHAT REMAINS GENUINELY UNKNOWN

Longitudinal evidence on the effects of different assessment regimes — over years, not weeks — does not exist. The recommendation that assessment design should attend to motivation, self-regulation, and character development over multi-year educational trajectories is based on theoretical convergence, not on longitudinal data.

Whether any assessment system can fully satisfy both the formative-learning function and the institutional- accountability function without compromise remains an open question. The programmatic-assessment model from medical education is the closest available answer, and it requires institutional investment that most educational contexts cannot match.

Whether the cross-cultural evidence (Section X) generalizes — whether the assessment-motivation findings from Western SDT research hold in all educational cultures — is partially addressed but not resolved. The Chinese and Korean evidence suggests the tension is cross-cultural; the Finnish evidence suggests alternative institutional designs are possible; but the causal mechanisms underlying cultural variation remain underspecified.

The ill-structured-domain feedback guidance in Section VII is based on convergent practitioner traditions rather than experimental evidence. The cross-walk between these traditions and the cognitive-psychology frameworks is this review's contribution; the experimental confirmation that the convergent principles produce better outcomes than well-structured-domain feedback methods in ill-structured domains does not exist.

The AI-feedback literature is 24 months old. Every finding in Section IX is provisional. The metacognitive-laziness concern is well-documented but the interventions that might mitigate it are theoretically motivated, not empirically validated.

12.4 WHAT THIS MEANS FOR THE LAB

For the Applied Pedagogy curriculum, the evidence supports a specific assessment architecture — the three-tier system described in Section XI — with domain-specific and learner-level-specific feedback design. The architecture is evidence-based where the evidence exists, principled where it does not, and honest about the difference. The assessment- motivation tension is not solved but is narrowed to specific design choices that the evidence can inform: ungraded formative practice, criterion-referenced evaluation, feedback separated from grades, self- and peer assessment developed as central practices, AI used as a task-level supplement with metacognitive safeguards.

What the lab still needs, and what this review cannot provide, is the longitudinal evidence that would confirm whether these design choices produce the multi-year motivational and developmental outcomes the theoretical convergence predicts. That evidence does not exist for any assessment system. The best the lab can do is design well, monitor honestly, and revise when the evidence warrants it.

Review complete. Word count: approximately 22,000 words.

REFERENCES

Banihashem, S. K. et al. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education*, 21, 37.

Bauer, E. et al. (2025). Looking beyond the hype: Understanding the effects of AI on learning outcomes. *British Journal of Educational Technology*.

Bego, C. R. et al. (2024). Nine-course study of retrieval practice in STEM. *Journal of Educational Psychology*.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about Knowing* (pp. 185–205). MIT Press.

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

Black, P. & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, 25(6), 545–567.

Blum, S. D. (ed.) (2020). *Ungrading: Why Rating Students Undermines Learning (and What to Do Instead)*. West Virginia University Press.

Bureau, J. S. et al. (2021). Pathways to student motivation: A meta-analysis of antecedents of autonomous and controlled motivations. *Review of Educational Research*, 92(4), 527–578.

Butler, R. (1988). Enhancing and undermining intrinsic motivation: The effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of Educational Psychology*, 58(1), 1–14.

Butler, A. C. & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, 36(3), 604–616.

Cameron, J. (2001). Negative effects of reward on intrinsic motivation — a limited phenomenon: Comment on Deci, Koestner, and Ryan (2001). *Review of Educational Research*, 71(1), 29–42.

Carless, D. & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8), 1315–1325.

Carless, D. & Chan, K. K. H. (2016). Managing dialogic use of exemplars. *Assessment & Evaluation in Higher Education*, 42(6), 930–941.

ten Cate, O. (2013). Nuts and bolts of entrustable professional activities. *Journal of Graduate Medical Education*, 5(1), 157–158.

ten Cate, O. & Regehr, G. (2018). The power of subjectivity in the assessment of medical trainees. *Academic Medicine*, 93(11), 1584–1585.

Chung, J.-H. & Park, Y.-S. (2024). Academic stress and mental health among high school students in South Korea. *Asian Journal of Psychiatry*, 93, 103917.

Danielewicz, J. & Elbow, P. (2009). A unilateral grading contract to improve learning and teaching. *College English*, 72(3), 244–268.

Darvishi, A. et al. (2023). Impact of AI assistance on student agency in peer assessment. *Computers & Education*, 206, 104905.

Deci, E. L., Koestner, R. & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6), 627–668.

Eisner, E. W. (1976). Educational connoisseurship and criticism: Their form and functions in educational evaluation. *Journal of Aesthetic Education*, 10(3/4), 135–150.

Escalante, J. et al. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student perceptions. *International Journal of Educational Technology in Higher Education*, 20(1), 57.

Eva, K. W. & Regehr, G. (2005). Self-assessment in the health professions: A reformulation and research agenda. *Academic Medicine*, 80(10), S46–S54.

Fan, Y. et al. (2024). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, 55(3), 1186–1205.

Fournier, J. P., Demeester, A. & Charlin, B. (2008). Script concordance tests: Guidelines for construction. *BMC Medical Education*, 8(1), 31.

Garon-Carrier, G. et al. (2016). Intrinsic motivation and achievement in mathematics in elementary school: A longitudinal investigation of their association. *Child Development*, 87(1), 165–175.

Van Gog, T. & Sweller, J. (2015). Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review*, 27(2), 247–264.

Goldschmidt, G., Hochman, H. & Dafni, I. (2010). The design studio "crit": Teacher-student communication. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 24(3), 285–302.

Guo, K. et al. (2024). Effects of AI-supported approach to peer feedback on pre-service teachers' online peer feedback quality. *Computers & Education*, 214, 104961.

Harden, R. M. (1999). AMEE Guide No. 14: Outcome-based education: Part 1. *Medical Teacher*, 21(1), 7–14.

Harden, R. M. (2016). Revisiting 'Assessment of clinical competence using an objective structured clinical examination (OSCE).' *Medical Education*, 50(4), 376–379.

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

Hendrickson, K. A. (2012). Assessment in Finland: A scholarly reflection on one country's use of formative, summative, and evaluative practices. *Mid-Western Educational Researcher*, 25(1/2), 33–43.

Kasneci, E. et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274.

Kingston, N. & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, 30(4), 28–37.

Kirkpatrick, R. & Zang, Y. (2011). The negative influences of exam-oriented education on Chinese high school students: Backwash from classroom to child. *Language Testing in Asia*, 1(3), 36.

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

Klieme, E. et al. (2003). *Zur Entwicklung nationaler Bildungsstandards: Eine Expertise.* BMBF. [German]

Kluger, A. N. & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284.

Koretz, D. et al. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice*, 13(3), 5–16.

Link, L. J. & Guskey, T. R. (2022). Is standards-based grading effective? A mixed methods systematic review of academic research. *Theory Into Practice*, 61(2), 127–137.

Liu, G. X. Y. & Helwig, C. C. (2020). Autonomy, social inequality, and support in Chinese urban and rural adolescents' reasoning about the gaokao exam system. *British Journal of Developmental Psychology*, 39(1), 106–122.

Lubarsky, S. et al. (2011). Script concordance testing: A review of published validity evidence. *Medical Education*, 45(4), 329–338.

Lubarsky, S. et al. (2015). Using script theory to cultivate illness script formation and clinical reasoning in health professions education. *Canadian Medical Education Journal*, 6(2), e61.

Marzano, R. J. (2000). *Transforming Classroom Grading.* ASCD.

Meyer, J. et al. (2023). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers & Education: Artificial Intelligence*, 4, 100121.

Mullet, H. G. et al. (2014). Delaying feedback promotes transfer of knowledge despite student preferences to receive feedback immediately. *Journal of Applied Research in Memory and Cognition*, 3(3), 222–229.

Nicol, D. (2020). The power of internal feedback: Exploiting natural comparison processes. *Assessment & Evaluation in Higher Education*, 46(5), 756–778.

Nieminen, J. H. & Carless, D. (2022). Feedback literacy: A critical review of an emerging concept. *Assessment & Evaluation in Higher Education*, 48(1), 21–34.

North, S. (1987). *The Making of Knowledge in Composition.* Boynton/Cook.

O'Connor, K. (2010). *A Repair Kit for Grading: 15 Fixes for Broken Grades.* 2nd ed. Pearson.

Panadero, E. & Lipnevich, A. (2021). A review of feedback models and typologies: Towards an integrative model of feedback elements. *Educational Research Review*, 35, 100416.

Panadero, E., Jönsson, A. & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review*, 22, 74–98.

Poehner, M. E. & Lantolf, J. P. (2005). Dynamic assessment in the language classroom. *Language Teaching Research*, 9(3), 233–265.

Prentice, S. et al. (2020). Workplace-based assessment failure modes. *Medical Education*, 54(11), 1016–1024.

Rheinberg, F. (2001). Bezugsnormen und schulische Leistungsbeurteilung. In F. E. Weinert (Ed.), *Leistungsmessungen in Schulen*. Beltz. [German]

Roediger, H. L. & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249–255.

Ryan, R. M. & Weinstein, N. (2009). Undermining quality teaching and learning: A self-determination theory perspective on high-stakes testing. *Theory and Research in Education*, 7(2), 224–233.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119–144.

Sadler, D. R. (2010). Beyond feedback: Developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education*, 35(5), 535–550.

Sadler, D. R. (2008). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education*, 33(2), 159–179.

Schimmer, T. (2016). *Grading from the Inside Out.* Solution Tree.

Sorensen-Unruh, C. (2024). The ungrading learning theory we have is not the ungrading learning theory we need. *CBE—Life Sciences Education*, 23(2), es4.

Sortwell, A. et al. (2024). A systematic review of meta-analyses on the impact of formative assessment on K-12 students' learning. *Education Sciences*, 14(3), 306.

Tai, J. et al. (2017). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education*, 76, 467–481.

To, J., Panadero, E. & Carless, D. (2021). A systematic review of the educational uses and effects of exemplars. *Assessment & Evaluation in Higher Education*, 47(8), 1209–1228.

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

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

Wormeli, R. (2018). *Fair Isn't Always Equal: Assessment and Grading in the Differentiated Classroom.* 2nd ed. Stenhouse.

Yan, L. et al. (2023). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(4), 1377–1399.

Zhan, Y. & Yan, Z. (2025). Students' engagement with ChatGPT feedback: Implications for feedback literacy development. *Assessment & Evaluation in Higher Education*.