HOW SHOULD INSTRUCTION BE DESIGNED?

From Binary Debate to Expertise-Adaptive Design Across Domains

Applied Pedagogy Research Lab

Guido Bartolucci, Principal Investigator

guido@appliedpedagogy.com

LAB.APPLIEDPEDAGOGY.COM

W2-004 · April 2026

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

CONTENTS

1

THE INSTRUCTIONAL-DESIGN QUESTION, REFRAMED

How should a teacher design instruction? For two decades the field has treated this as a binary: direct instruction or inquiry? The framing was crystallized in 2006 when Paul Kirschner, John Sweller, and Richard Clark published "Why Minimal Guidance During Instruction Does Not Work," arguing that constructivist, discovery-based, and inquiry-based approaches all share a fatal flaw — they overload the limited capacity of working memory by asking novice learners to simultaneously discover solutions and learn from the discovery process (Kirschner, Sweller & Clark, 2006)[○]. Within a year, Cindy Hmelo-Silver, Ravit Golan Duncan, and Clark Chinn responded that the argument rested on a mischaracterization: well-designed problem-based and inquiry learning are not "minimally guided" at all but involve substantial scaffolding, structured problems, and expert facilitation (Hmelo-Silver, Duncan & Chinn, 2007)[○].

Nearly two decades later, this review argues that the binary is not just unhelpful but actively misleading. The evidence does not support the superiority of either approach in isolation. What it supports is an *expertise-adaptive model of instruction* in which the design of instruction should change as learners develop from novice to expert — and should change differently depending on whether the domain is well-structured or ill-structured. The real questions are not "which approach wins?" but "what should instruction look like at each stage of learning, what signals the transitions between stages, and how does the design change when the domain resists the clean decomposition that cognitive load theory assumes?"

This review is a refinement of the L1-004 investigation, which established the broad outlines of the expertise-adaptive argument but left critical gaps. v1 described the worked-example-to-inquiry trajectory in the abstract but did not specify the stages concretely — what instruction actually looks like at each point, what diagnostic markers signal readiness for the next stage, how the design changes when a teacher moves from teaching algebra to teaching essay writing. v1 flagged eight gaps; the most consequential was Gap 1, the near-total absence of guidance for instructional design in ill-structured domains — writing, design, ethical reasoning, clinical judgment, artistic practice. These are not marginal cases. Applied Pedagogy's curriculum mission — building competence across all five layers of the competence stack — requires instruction that develops judgment, metacognition, and character, all of which are ill-structured. A theory of instructional design that only works for mathematics is not a theory of instructional design; it is a theory of mathematics instruction. v1 engaged the primary frameworks (Rosenshine, Merrill, Chi's ICAP, Engelmann's Direct Instruction, van Merriënboer's 4C/ID) at secondary-source depth. v2 reads the primary texts.

Three developments since v1 reshape the investigation. First, W2-008's review of curriculum philosophy concluded that the capacity for warm, reciprocal relationship is the single strongest predictor of adult flourishing, and that self-regulation develops primarily through warm, predictable environments rather than direct cognitive training (Watts et al., 2018, as reported in W2-008 §3.3). This does not overturn the case for explicit instruction, but it reframes it: an instructional sequence that is cognitively optimal but destroys the relational-environmental conditions for self-regulation is not, on balance, a good sequence. Instructional design cannot treat the relational context as someone else's problem.

Second, W2-009's review of competence formation described the skill-to-judgment transition — the Layer 2→3 shift in the lab's competence stack — as a qualitative reorganization of knowledge

from surface features to deep structure (Chi, Feltovich & Glaser, 1981, as discussed in W2-009 §II), accompanied by the progressive transfer of control from explicit analytical processing to tacit, intuitive pattern recognition (Polanyi, 1966; Klein, 1998, as discussed in W2-009 §III). This transition cannot be directly taught; it must be developed through accumulated experience with varied, consequential situations in environments that provide valid feedback. The instructional-design question is therefore: what sequence of designed experiences builds the representations that support judgment? Worked examples can build skill, but what builds judgment?

Third, the AI-mediated learning environment has arrived. Fan et al. (2024) found that ChatGPT-assisted learners improved task performance but showed no gains in knowledge or transfer — a phenomenon the authors term "metacognitive laziness" (Abstract-verified). This finding connects directly to Brousseau's fundamental paradox of instruction: "everything that [the teacher] undertakes in order to make the student produce the behaviours that she expects tends to deprive this student of the necessary conditions for [...] learning" (Brousseau, 1997, §1.2.2)•. AI tutoring without instructional-design principles risks the Topaze effect at industrial scale — scaffolding that eliminates the learning target itself.

This review proceeds in twelve sections. We begin by establishing the cognitive-science foundation — not repeating what W2-001 covers in depth but identifying the specific implications for instructional design. We then trace the inquiry debate, productive failure, the 4C/ID model, and the practitioner frameworks (Rosenshine, Merrill, Chi, Engelmann), engaging each at primary-text depth. The central synthetic contribution is Section 7: a concrete, five-stage expertise-adaptive model with transition signals, worked examples for both well-structured and ill-structured domains, and explicit documentation of where the evidence is strong and where it runs thin. Section 8 addresses the ill-structured domain gap — the most important gap in v1 and the most consequential for Applied Pedagogy's curriculum mission — through direct engagement with writing pedagogy (Hillocks, 1986), design education (Schön, 1983, 1987), French *didactique* (Brousseau, 1997; Chevallard, 1985), Russian activity theory (Galperin, via Engeness, 2020), German *Didaktik* (Klafki, via Sjöström & Eilks, 2020), and Japanese lesson study (Takahashi & McDougal, 2016). Sections 9–10 address the autonomy-structure tension and AI-mediated design. Section 11 translates findings into specific prescriptions for curriculum design. Section 12 provides an honest closing assessment.

Throughout, every claim carries a provenance tag — Verified (direct), Verified (via PI summary), Abstract-verified, or Training-derived — so the reader can calibrate trust claim by claim. The provenance breakdown is reported in 'changelog.md'.

2 THE COGNITIVE-SCIENCE FOUNDATION

This section establishes the cognitive-science constraints on instructional design. W2-001 provides the full treatment of cognitive architecture, working memory, and schema theory; here we address only what is specific to the design of instruction.

2.1 WORKING MEMORY AND THE CASE FOR EXPLICIT INSTRUCTION

The central constraint is simple: human working memory can process approximately four chunks of novel information simultaneously (Cowan, 2001)○. When the demand exceeds this capacity, learning fails — not as a failure of motivation or effort but as a structural limitation of cognitive architecture. Long-term memory has no known capacity limits; expertise consists of well-organized schemas stored there that allow complex patterns to be treated as single chunks. The transition from novice to expert is, at its cognitive core, the construction of increasingly sophisticated schemas that compress information and reduce the load on working memory.

Cognitive load theory (CLT) provides the framework for understanding how instruction interacts with this architecture. The 2019 retrospective by Sweller, van Merriënboer, and Paas grounded CLT in evolutionary psychology, distinguishing biologically primary knowledge — acquired effortlessly through evolution, such as spoken language and facial recognition — from biologically secondary knowledge — acquired only through explicit instruction, such as reading, mathematics, and scientific reasoning. The implication: "Instruction should be explicit because we have evolved to learn directly from other people via the borrowing and reorganising principle. [...] it needs to be organised in a manner that reduces working memory load because working memory load primarily occurs when processing novel, domain-specific information" (Sweller, van Merriënboer & Paas, 2019, p. 274)●.

This is the strongest version of the case for explicit instruction, and it is well-grounded. But CLT's own authors recognize its limits. The 2019 retrospective formally integrated the 4C/ID model as a "twin theory," acknowledging that "CLT alone is sufficient to develop a useful instructional design model for complex learning at the level of whole educational programs" is not the case — "to ensure that the freed-up resources are actually devoted to learning, the Ten Steps relies on several specific learning theories" (van Merriënboer & Kirschner, 2018, §0.8.42)●. And the 2019 paper opened a door CLT had previously kept closed: it acknowledged that for professional education, "emotions, stress and uncertainty may restrict the capacity of working memory" and that "educational programs must be carefully designed in such a way that learners develop professional competencies enabling them to perform professional tasks up to the standards, including the ability to deal with emotions, stress and uncertainty" (Sweller et al., 2019, p. 285)●. This is CLT's closest approach to the ill-structured domains where emotions and uncertainty are not extraneous load but part of the task itself.

2.2 THE WORKED-EXAMPLE EFFECT AND ITS BOUNDARY CONDITION

The worked-example effect — novice learners who study step-by-step demonstrations learn more efficiently than novices who attempt equivalent problems on their own — is one of CLT's most

robust findings, replicated across mathematics, physics, computer programming, and other well-structured domains (Training-derived). The mechanism is straightforward: worked examples free working memory from the demands of means-ends analysis (searching for a solution path) and redirect it toward the structure of the solution — the underlying principles and the relationships between steps. This structural attention builds schemas.

The worked-example effect has a crucial boundary condition. Kalyuga, Ayres, Chandler, and Sweller (2003) demonstrated the *expertise reversal effect*: instructional techniques optimal for novices become ineffective or harmful as learners gain expertise (Training-derived). The mechanism: as learners develop schemas, guidance that was once helpful becomes redundant with their existing knowledge, and processing redundant information imposes extraneous load. Sweller et al. (2019) reconceptualized the expertise reversal effect as a "compound effect" — "a variant of the more general element interactivity effect. [...] Instructional procedures designed for novices dealing with multiple, interacting elements can be counterproductive as expertise increases and the interacting elements become embedded in knowledge structures held in long-term memory" (p. 277)●.

Tetzlaff et al. (2025) provided the first comprehensive meta-analysis of the expertise reversal effect — 60 studies, 176 effect sizes, n = 5,924 (Abstract-verified). The findings confirm the effect's robustness but add important boundary conditions: low-prior-knowledge learners benefit from high-assistance instruction (d = 0.505), high-prior-knowledge learners benefit from low-assistance instruction (d = −0.428), and the effect is asymmetric — providing novices with assistance has a *stronger* effect than withholding it from experts. Critically for this review, the effect is moderated by educational status and content domain: "for humanities and language learning, the evidence for effectiveness is less clear" (Tetzlaff et al., 2025)○. This domain boundary condition connects directly to the ill-structured domain gap — worked examples may be less effective in domains where "correct answers" cannot be specified.

The practical implication is the *guidance-fading effect*: "For novices, additional information or particular activities such as studying worked examples may be essential. With increases in expertise, these same activities may become redundant and impose an unnecessary cognitive load. Past a certain point, studying worked examples may be counterproductive and they should be faded out and replaced by problems" (Sweller et al., 2019, p. 277)●. This is the CLT- internal mechanism for the instructional transition from explicit instruction to independent practice — and, as we shall see, the foundation for the staged-instruction model this review proposes.

2.3 THE GUIDANCE-FADING EFFECT AND ITS DESIGN IMPLICATIONS

The practical consequence of the expertise reversal effect is the *guidance-fading effect*. Sweller et al. (2019) state the principle clearly: "For novices, additional information or particular activities such as studying worked examples may be essential. With increases in expertise, these same activities may become redundant and impose an unnecessary cognitive load. Past a certain point, studying worked examples may be counterproductive and they should be faded out and replaced by problems" (p. 277)●. This is the CLT-internal mechanism for the instructional transition from explicit instruction to independent practice.

Renkl and Atkinson (2003) operationalized this with the fading-guidance strategy: worked examples are gradually transformed into practice problems by successively removing steps, forcing the learner to complete increasingly large portions of the solution independently (Training-derived). The 4C/ID model implements the same principle through its completion strategy — "learners

first study cases, then work on completion tasks, and finally perform conventional tasks" (van Merriënboer & Kirschner, 2018, §0.10.47)•.

The design challenge is *when* to fade. The guidance-fading effect tells us that fading should occur as expertise develops, but it does not specify the diagnostic markers. Two rough heuristics exist: Rosenshine's 80% success-rate threshold (from one study, as noted above) and van Merriënboer's "horizontal standards matrix" — a framework for specifying performance standards within each task class. Neither provides the real-time, learner-specific diagnostic that adaptive instruction requires. Sherrington (2019) honestly acknowledges this: "judging the transition from students being guided enough to becoming independent is a subtle skill, a central element of teacher expertise" (Strand 4)•. The transition signal for adaptive instruction remains the central unsolved problem — see Section 7 for the staged model's treatment and its honest limitations.

2.4 WHAT CLT DOES NOT ADDRESS

Three domains central to instructional design fall outside CLT's explanatory scope.

First, CLT does not address *problem-setting* — the construction of a problem from ambiguous materials. All CLT- derived instructional techniques (worked examples, completion problems, goal-free problems) assume the problem is given. Schön (1983) argued that "with this emphasis on problem solving, we ignore problem setting, the process by which we define the decision to be made, the ends to be achieved, the means which may be chosen" (§1.4.38)•. In ill-structured domains — design, clinical reasoning, policy analysis — the most consequential professional work is in constructing the problem, not solving it. CLT has nothing to say about how to teach this.

Second, CLT does not address the *relational context* of instruction. Whether a worked example is delivered in a warm, autonomy-supportive environment or a cold, controlling one is invisible to CLT's analysis. Yet W2-008 established that the relational-environmental conditions are preconditions for the self-regulation that instruction ultimately aims to develop. An instructional sequence that is CLT-optimal but relationally destructive is not, on balance, well-designed.

Third, CLT does not address the *social-constructive* dimension of learning. Vygotsky's zone of proximal development, Galperin's stepwise formation of mental actions, and Brousseau's theory of didactical situations all theorize learning as inherently social in ways CLT's information- processing framework does not capture. CLT treats social interaction as a context variable; these traditions treat it as constitutive of learning itself.

These are not criticisms of CLT — a theory of cognitive load need not address everything. They are scope conditions. The instructional designer needs CLT's constraints *and* the frameworks that address what CLT cannot.

3 THE INQUIRY AND PROBLEM-BASED LEARNING CASE

3.1 WHAT HMELO-SILVER ET AL. ACTUALLY ARGUED

Hmelo-Silver, Duncan, and Chinn (2007) did not argue that inquiry is always better than direct instruction. They argued that Kirschner et al. "have mistakenly conflated PBL and IL with discovery learning," and that well-designed PBL and inquiry "employ scaffolding extensively thereby reducing the cognitive load and allowing students to learn in complex domains" (Hmelo-Silver et al., 2007)[○]. The distinction is between unguided discovery (which the evidence shows is ineffective for novices) and scaffolded inquiry (which the evidence shows can be highly effective). The question is not "guided versus unguided" but "what kind of guidance, and how much, at what point in learning?"

This rebuttal is largely convincing. The PBL literature is clear that effective PBL requires extensive design — structured problems, skilled facilitation, and calibrated scaffolding. When PBL fails, it typically fails because one or more of these design elements is absent. Hmelo-Silver (2004) showed that the approach works by activating prior knowledge, supporting knowledge construction through collaborative problem-solving, and developing self-directed learning skills — all mechanisms consistent with CLT's emphasis on schema construction (Training-derived).

3.2 THE META-ANALYTIC EVIDENCE

The meta-analytic literature reveals a consistent pattern that neither side of the binary debate fully acknowledges.

Lazonder and Harmsen (2016) — the most methodologically rigorous meta-analysis of inquiry-based learning (72 studies, FWCI 209.59) — found a substantial overall effect of guidance on learning outcomes ($d = 0.50$) and that guided inquiry outperformed unguided inquiry across all outcome measures (Training-derived). This finding is important for what it does and does not say. It says that guidance matters enormously — confirming CLT's central claim that novices need support. But it also says that *guided* inquiry is highly effective, which challenges any reading of CLT that treats inquiry itself as the problem. The type of guidance moderated effects on performance success but not on learning outcomes, suggesting that many different forms of scaffolding can work — the key is that scaffolding is present.

The PBL meta-analyses tell a more nuanced story than either camp typically acknowledges. Dochy et al. (2003) — the earliest major PBL meta-analysis, primarily from medical education, FWCI 125.54 — found robust positive effects on skills and application but less consistent effects on knowledge as measured by conventional tests (Training-derived). Strobel and van Barneveld (2009) synthesized multiple meta-analyses and found PBL superior for long-term retention, skill development, and student satisfaction, but traditional instruction more effective for short-term retention on standardized exams (Training-derived). Walker and Leary (2009) provided the most granular analysis, finding that PBL effects were larger for application and problem-solving assessments than for factual recall, and larger in medical education than other disciplines (Training-derived).

The convergent finding across these meta-analyses is striking: PBL tends to produce equivalent or slightly lower scores on immediate factual recall tests but equal or superior performance on assessments of application, clinical reasoning, problem-solving skill, and long-term retention. The

outcome measure determines which approach "wins." Studies that measure factual recall favor direct instruction; studies that measure conceptual understanding, application, and long-term retention favor scaffolded inquiry. This is not a minor methodological point — it means that the entire debate has been confounded by what researchers chose to measure.

Alfieri et al. (2010) provided the clearest quantitative summary with a two-part meta-analysis: unassisted discovery learning failed relative to explicit instruction ($d = -0.38$), but enhanced or guided discovery outperformed other forms of instruction ($d = 0.30$) (Training-derived). The take-away: discovery *per se* is not the mechanism — the mechanism is active generation within a structured context.

Freeman et al. (2014) conducted a landmark meta-analysis of 225 studies showing that "active learning" in STEM increases performance by $g = 0.47$ and reduces failure rates by 55% compared to traditional lecturing (Verified direct). The failure- rate finding is particularly striking: "if the experiments analyzed here had been conducted as randomized controlled trials of medical interventions, they may have been stopped for benefit — meaning that enrolling students in traditional lecture courses is unethical in light of the evidence that active learning is more effective" (Freeman et al., 2014)●. However, this finding must be interpreted with care: "active learning" as defined by Freeman et al. is so broad that it includes everything from clicker questions during lectures to full studio redesigns — and explicit instruction with guided practice (Rosenshine, Engelmann) qualifies as "active learning" under their definition. The finding is against passive lecturing, not against explicit instruction.

3.3 THE DE JONG–SWELLER CONVERGENCE

The most telling development in the debate came in 2023–2024. Thirteen leading inquiry-learning researchers — including de Jong, Hmelo-Silver, Koedinger, and Linn — jointly argued that "a combination of inquiry and direct instruction may often be the best approach to support student learning" (de Jong et al., 2023)○. Sweller's response acknowledged "potential agreement with De Jong et al. on the essential role of explicit instruction" while arguing that the expertise reversal effect already provides the framework for when to combine the two approaches (Sweller, 2023)○.

The debate has converged. Both sides now agree that explicit instruction is essential and that the question is sequencing and combination — not binary choice. The staged-instruction model proposed in Section 7 is consistent with both positions.

3.4 WHAT THE META-ANALYSES CONVERGE ON

Five findings are robust across the literature:

1. **Unguided discovery is ineffective for novices.** Kirschner et al. were right about this. 2. **Guided inquiry is effective, often more effective than direct instruction for conceptual understanding.** 3. **Direct instruction is more efficient for procedural knowledge and short-term recall.** 4. **The outcome measure matters enormously.** Studies that measure factual recall favor direct instruction; studies that measure conceptual understanding and transfer favor guided inquiry. 5. **Prior knowledge is the critical moderator.** The expertise reversal effect operates at the level of instructional design.

The practical implication is that the binary framing was always the wrong question. The right question is what v1 identified and what this review attempts to answer concretely: how should instruction change across the arc of learning?

4

PRODUCTIVE FAILURE AS BRIDGE

4.1 KAPUR'S FRAMEWORK

Manu Kapur's program of research on productive failure provides the most important bridge between the direct- instruction and inquiry traditions. The core finding, replicated across more than 50 studies: inverting the standard instructional sequence — problem solving *before* instruction (PS-I) rather than instruction before problem solving (I-PS) — produces deeper conceptual understanding and transfer, even when initial problem-solving attempts fail (Kapur, 2024)•.

The distinction from discovery learning is critical. "Productive Failure is not the same as Discovery Learning, for it combines the benefits of constrained discovery learning through problem solving, followed by instruction" (Kapur, 2024, §7.5.3)•. PF is a two-phase design: struggle first, then explicit instruction. The instruction phase typically involves the teacher presenting canonical concepts and procedures — connecting them explicitly to the students' generated solutions.

4.2 THE 4A FRAMEWORK

Kapur identifies four mechanisms:

Activation. The problem-solving attempt activates learners' prior knowledge — even incomplete, incorrect prior knowledge. "This, however, results in a strong activation of the cognitive system because it activates relevant prior knowledge needed for learning" (§8.0.4)•.

Awareness. Attempting and failing creates metacognitive awareness of knowledge gaps — "a gap between what we know and what we need to know" (§8.0.4)•.

Affect. The struggle generates "needs, interest, curiosity, motivation, and emotions — to learn from subsequent instruction" (§8.0.4)•.

Assembly. The instruction phase "assembles" prior attempts into correct understanding. "An expert or a teacher can help assemble our knowledge into the correct concepts and ideas" (§8.0.6)•.

4.3 IS PRODUCTIVE FAILURE IN TENSION WITH CLT?

On the surface, PF appears to contradict CLT: if novices have limited working memory and worked examples are optimal, how can asking novices to solve unsolvable problems be beneficial? Kirschner et al. (2006) would seem to predict that the generation phase should overload working memory and produce inferior learning. The resolution lies in understanding PF as a two-phase design, not as unguided discovery. The struggle phase is not where learning occurs — it is where the *conditions* for learning are prepared. The instruction phase that follows is where schemas are built, and the prior struggle makes this phase more effective by activating relevant knowledge, creating awareness of gaps, and generating motivation to learn.

The critical distinction: PF is not minimally guided instruction. It is an instructional design with a specific structure (problem-solving THEN instruction), specific task requirements (challenging but accessible, multiple solution paths), and a specific mechanism (cognitive preparation, not discovery learning). The learner does not discover the answer through struggle; the learner prepares to understand the answer through struggle.

Kapur's design principles make the specificity clear. PF tasks must be "challenging but accessible" — students must possess enough prior knowledge to generate suboptimal but relevant solutions (§14.2.1)•. The tasks must invite "multiple solutions" and use "layperson's language" without technical jargon (§14.2.1, §14.4.1)•. The problem space must be "rich enough for learners to generate multiple representations and solution methods" (§14.4.1)•. And the instruction phase must explicitly connect the canonical solution to the students' generated approaches — the Assembly mechanism depends on the activation that preceded it. When these design conditions are met, "the relative effect of learning from Productive Failure was up to three times (that's 300%) that of learning from a good teacher for one year" (§7.12.9)•.

But there is an important boundary condition that Kapur acknowledges in design principle but does not fully theorize: PF requires that students possess *sufficient prior knowledge* to generate relevant (if incorrect) approaches. For genuine novices with no relevant schemas, the struggle phase may produce nothing useful. This is not a refutation of PF — it is a staging condition. PF is most powerful at the stage when learners have enough foundational knowledge (from prior explicit instruction) to engage productively with the problem but not yet enough to solve it correctly. The staged- instruction model places PF at Stage 3, after explicit instruction (Stage 1) and scaffolded practice (Stage 2) have built the schemas that productive activation requires.

4.4 LEARNING PATH DEPENDENCE

Kapur's most provocative claim: "What my results show is how one learns the foundational knowledge influences how well they understand it and can transfer it. The learning path matters" (§7.13.2)•. This directly challenges the assumption — implicit in much of the DI literature — that the only thing that matters is *what* is learned, not *how* it was learned. Two students may score identically on a knowledge test but differ profoundly in their ability to transfer that knowledge to new situations, depending on whether they acquired it through direct instruction alone (I-PS) or through productive failure followed by instruction (PS-I). The *sequence* of instruction is itself an instructional variable, and the variable matters most precisely where the curriculum cares most — at the level of conceptual understanding and transfer, not procedural recall.

The implication for curriculum design is substantial. If learning path matters, then curriculum cannot be designed solely by specifying *what* students should know — it must also specify *the sequence of learning experiences* through which they come to know it. Two curricula covering identical content may produce profoundly different outcomes depending on whether they sequence instruction as I-PS (which the traditional model assumes) or PS-I (which productive failure research recommends for conceptual topics after foundational knowledge is established). This finding elevates instructional design from a delivery question — "how do we communicate this content efficiently?" — to a learning-architecture question: "what sequence of experiences builds the deepest understanding?"

4.5 PRODUCTIVE FAILURE AND THE FRENCH *DIDACTIQUE*

A striking convergence exists between Kapur's productive failure and Brousseau's theory of didactical situations. Brousseau (1997) argued that "the student learns by adapting herself to a milieu which generates contradictions, difficulties and disequilibria" (§1.0.7)• — the French tradition's independent discovery of the productive-failure principle from a completely different theoretical base. Where Kapur's theoretical lineage runs through desirable difficulties and the generation effect, Brousseau's runs through Piaget and epistemological obstacles. The instructional design

principle is strikingly similar: a designed environment that generates productive cognitive conflict, followed by formalization.

The convergence goes deeper. Brousseau's concept of *devolution* — "the act by which the teacher makes the student accept the responsibility for an (adidactical) learning situation or for a problem" (§5.1.4)• — is not the same as scaffolding-fading. Fading removes support from a fixed problem. Devolution transfers *responsibility for the problem itself*. This distinction matters for the staged-instruction model: the transition from Stage 2 to Stage 3 is not simply "less scaffolding" but a qualitative shift in who owns the problem.

4.6 PRODUCTIVE FAILURE, W2-008, AND METACOGNITIVE LAZINESS

Fan et al.'s (2024) finding that AI-assisted learners improve task performance but not knowledge or transfer — "metacognitive laziness" — is the Topaze effect for the AI age. Brousseau defined the Topaze effect as what happens "by choosing easier and easier questions, the teacher tries to achieve the optimum meaning for the maximum number of students. If the target knowledge disappears completely, we have the Topaze effect" (§1.0.2)•. When an AI provides answers rather than designed struggle, the learning target disappears. The productive-failure framework provides the design alternative: AI should generate productive struggle and then scaffold assembly, not bypass the learning process.

5

THE 4C/ID MODEL AND COMPLEX LEARNING

5.1 THE MODEL AT PRIMARY-TEXT DEPTH

Van Merriënboer and Kirschner's Four-Component Instructional Design (4C/ID) model is the most architecturally sophisticated framework for instructional design in complex domains. Reading *Ten Steps to Complex Learning* (3rd ed., 2018) at primary-text depth reveals a framework more nuanced than secondary accounts suggest.

The model rests on a foundational distinction between *recurrent* and *non-recurrent* skills. "Constituent skills are classified as nonrecurrent skills if they need to be performed as schema-based processes after the training: These are the problem-solving, reasoning, and decision-making aspects of behavior. [...] Constituent skills are classified as recurrent skills if they will be performed as rule-based processes after the training, routine aspects and sometimes fully automatic aspects of behavior" (§0.8.36)•. "The classification of skills as nonrecurrent or recurrent is important in the Ten Steps because instructional methods for the effective and efficient acquisition of them are very different" (§0.8.37)•.

This distinction maps directly onto the well-structured/ ill-structured divide. Recurrent skills (solving quadratic equations, performing a blood draw, conjugating verbs) respond to procedural information — step-by-step, just-in-time instruction that is faded as automaticity develops. Non-recurrent skills (diagnosing a patient, designing a building, writing a persuasive essay) require *supportive information* — domain models and cognitive strategies presented before or during a task class, not faded but elaborated as complexity increases.

5.2 THE FOUR COMPONENTS

Learning tasks are whole, authentic tasks organized in "task classes" of increasing complexity, with support fading within each class. The key design feature is the *completion strategy*: "learners first study cases, then work on completion tasks, and finally perform conventional tasks. [...] This completion strategy has been applied in several other domains, and experimental studies carried out there have consistently shown positive effects on learning and transfer" (§0.10.47)•. This is the 4C/ID version of the worked- example-to-fading transition.

Supportive information serves non-recurrent skills — "mental models and cognitive strategies that help learners perform the non-routine aspects of tasks" (§0.8.37, paraphrase from Verified direct). This is presented before the learner begins a new task class, not just-in-time.

Procedural information serves recurrent skills — "how-to instructions [...] best presented to learners exactly when they first need it to perform a task (i.e., just in time), after which it is faded for subsequent learning tasks as the learner masters it" (§0.8.38)•.

Part-task practice provides additional practice for sub-skills requiring automatization. While the model insists on whole-task practice as the primary vehicle, it acknowledges that some component skills need dedicated drill — calculation fluency in statistics, suturing technique in surgery, keyboard skills in programming. The design challenge is identifying which sub-skills truly require automated execution (recurrent skills) and dedicating practice time to them without fragmenting the whole task.

To illustrate: consider designing instruction for clinical diagnosis. The *learning task* is a whole case — "this patient presents with these symptoms; diagnose and recommend treatment." Task classes increase in complexity: Class 1 might be single- system presentations with classic symptoms; Class 3 might be multi-system presentations with ambiguous findings. Within each class, early cases come with heavy support (a completed diagnostic reasoning trace — the 4C/ID equivalent of a worked example) that fades across cases. *Supportive information* — mental models of disease mechanisms, cognitive strategies for differential diagnosis — is presented before the learner begins each new task class. *Procedural information* — how to conduct a focused physical examination, how to read a lab report — is provided just-in-time when first needed and then faded. *Part-task practice* targets recurrent skills needing automatization: reading EKGs, recognizing dermatological patterns.

This whole-task-first architecture is what makes 4C/ID distinct from — and in important ways superior to — the decomposition- first approach that dominates traditional instruction. Most teaching begins with isolated sub-skills and assumes that competence will emerge when they are combined. Van Merriënboer argues this assumption is wrong — the whole-task context is what gives the sub-skills meaning and motivational salience.

5.3 THE TRANSFER PARADOX

Van Merriënboer identifies a key challenge: "the methods that work the best for reaching isolated, specific objectives are often not the methods that work best for reaching integrated objectives and increasing transfer of learning" — the "transfer paradox" (§0.8.21)•. This is why the model insists on whole-task practice from the beginning: teaching sub-skills in isolation may build local competence efficiently but fails to develop the integrated performance that complex tasks demand.

5.4 IMPLEMENTATION CHALLENGES

The 4C/ID model's sophistication is also its weakness. "When teachers or novice designers use the Ten Steps for the first time, they often find the model difficult to apply" (§0.40.14)•, and "teacher design teams are strongly preferred above individual teachers designing their own learning tasks" (§0.40.16)•. The model was developed primarily for professional education — medicine, engineering, teacher training — where design resources are available. K-12 implementation evidence remains thin (v1 Gap 6). Frèrejean et al. (2019) documented implementation challenges even in higher education (Abstract-verified).

5.5 THE MODEL'S BLIND SPOT

The 4C/ID model acknowledges ill-structured tasks — "ill- structured problems [...] confront the task performer with unknown elements, have multiple acceptable solutions (or even no solution at all!), possess multiple criteria for evaluating solutions, and often require learners to make judgments" (§0.10.20)•. But it treats them primarily through the non-recurrent skill pathway — supportive information, cognitive strategies, mental models. It does not engage the specific pedagogical traditions (writing instruction, design education, clinical reasoning) that have developed domain-specific approaches to ill-structured instruction. The model is a powerful architecture that needs to be populated with domain-specific content the 4C/ID literature does not itself provide.

5.6 SELF-DIRECTED LEARNING AND SECOND-ORDER SCAFFOLDING

The model's treatment of the autonomy question is more nuanced than CLT's. Van Merriënboer introduces "second-order scaffolding" — "shared control over task selection where the learner and teacher/system work together to plan an optimal individualized learning trajectory" (§0.21.1)•. He acknowledges "a delicate balance between, on the one hand, autonomy, and, on the other hand, support and guidance. This balance needs to be carefully maintained in a process of second-order scaffolding" (§0.40.23)•. This is the closest the CLT-aligned literature comes to engaging the autonomy-structure tension as a design problem rather than a philosophical disagreement.

6 PRACTITIONER FRAMEWORKS AND THEIR EVIDENCE BASES

6.1 ROSENSHINE'S PRINCIPLES OF INSTRUCTION

Barak Rosenshine's (2012) "Principles of Instruction" in *American Educator* is arguably the most influential practitioner-oriented synthesis of instructional research. The article distills decades of research into ten principles — begin with review, present in small steps, ask many questions, provide models, guide practice, check for understanding, obtain a high success rate, scaffold difficult tasks, require independent practice, engage in regular review.

Reading the article in full (Verified direct) reveals several things the secondary accounts miss. First, Rosenshine grounds the principles in a "triple convergence" of three independent research traditions: "Even though these are three very different bodies of research, there is no conflict at all between the instructional suggestions that come from each of these three sources" (p. 12)•. The sources are cognitive science (CLT, worked examples), observational studies of "master teachers" (process-product research), and cognitive supports (scaffolding, reciprocal teaching). The methodological power of this convergence — three traditions, three methods, same conclusions — is genuinely impressive.

Second, Rosenshine's evidence base is almost entirely 1970s–1990s elementary mathematics and reading — well-structured domains with clear assessment criteria. The "master teachers" were identified by student achievement-test gains, which favor near transfer and procedural accuracy. The 80% success-rate threshold — "the optimal success rate for fostering student achievement appears to be about 80 percent" (p. 17)• — comes from a single fourth-grade math study. Its generalizability is unestablished.

Third, and most consequential for this review, Rosenshine writes that effective teachers "always did the experiential activities *after*, not before, the basic material was learned" (p. 12, emphasis added)•. This is precisely the claim productive failure challenges. Kapur's PS-I framework argues that exploration *before* instruction produces deeper learning. But Rosenshine's "after, not before" is an empirical generalization from master-teacher observation — not an experimental finding comparing sequencing options. The master teachers were doing what produced high scores on knowledge tests; productive failure's advantage appears on transfer measures the master-teacher studies did not use.

The honest assessment: Rosenshine's principles are the best available operationalization of Stage 1 instruction — what good explicit instruction looks like for novices in well-structured domains. They are not a complete instructional model. They say nothing about when to reduce scaffolding, when to introduce productive failure, or how to handle ill-structured content. This is not a criticism; it is a scope condition that Sherrington (2019) acknowledges in his practitioner translation: the principles apply to "well-structured" subjects (Sherrington, 2019, Introduction)•.

6.2 MERRILL'S FIRST PRINCIPLES

David Merrill's (2002) five "First Principles" — problem- centered, activation, demonstration, application, integration — attempt to identify features common to all effective instruction regardless of theoretical orientation (Training-derived). The framework is notable for its eclecticism: it

places both demonstration (explicit instruction) and application (problem-solving) within a single learning cycle, implicitly rejecting the binary framing.

The *activation* principle — "learning is promoted when learners activate relevant prior knowledge or experience" — aligns directly with productive failure research. Kapur's 4A framework begins with Activation for the same reason: activating prior knowledge before instruction enhances learning. Merrill arrived at this principle through a different analytical route — reviewing commonalities across multiple instructional design theories — but the convergence is telling. The *problem- centered* principle — "learning is promoted when learners engage in a task-centered instructional strategy" — is consistent with PBL, with 4C/ID's whole-task approach, and with the environmental mode. The *demonstration-application* sequence maps onto CLT's worked-example-to-practice transition. And the *integration* principle — "learning is promoted when learners integrate their new knowledge into their everyday world" — is one of the few instructional-design principles that addresses the transfer problem directly.

The limitation: Merrill's framework operates at a high level of abstraction. It tells you what features to include but not how to calibrate them to learner expertise, domain characteristics, or task structure. A Stage 1 lesson and a Stage 4 lesson could both satisfy all five principles while looking entirely different in practice. The evidence base for the five principles as a unified system is weaker than the evidence for each principle individually — the synthesis is more theoretical than empirical. Nevertheless, the First Principles serve a useful role as a design checklist: any instructional sequence that lacks activation, demonstration, application, or integration is probably missing something important.

6.3 CHI'S ICAP FRAMEWORK

Michelene Chi and Ruth Wylie's (2014) ICAP framework provides a taxonomy of learning activities by cognitive engagement: Passive (receiving without processing), Active (manipulating — highlighting, copying), Constructive (generating new information — self-explaining, creating analogies), and Interactive (collaborating to construct — building on each other's ideas in substantive dialogue) (Abstract-verified).

The framework predicts — and evidence supports — a hierarchy: Interactive > Constructive > Active > Passive for learning outcomes. The practical power lies in its independence from the direct-instruction-vs-inquiry framing: a lecture in which students generate self-explanations (Constructive) is more effective than a lecture with verbatim note-taking (Active), regardless of whether the instruction is "direct" or "inquiry-based." The critical variable is not who initiates the activity but what cognitive processing it demands.

Boundary conditions need acknowledgment. First, the hierarchy may interact with expertise level: a Constructive self- explanation task may overwhelm novices whose working memory is already taxed by the content itself. Asking a genuine novice to "explain why this solution works" when they cannot yet follow the solution is not Constructive learning — it is cognitive overload. The ICAP hierarchy implicitly assumes enough prior knowledge to engage at each level, making it complementary to, not in conflict with, CLT's guidance-fading prescription. For genuine novices, Active engagement (following along, highlighting key steps) may be the appropriate ceiling; Constructive engagement becomes possible as schemas develop.

Second, the conditions under which group interaction produces genuine Interactive learning — rather than social loafing, pooling of ignorance, or one student doing the work while others watch — are not fully specified. The ICAP framework predicts that Interactive > Constructive, but this prediction holds only when the interaction involves genuine co- construction — students

building on each other's ideas in substantive ways. Much of what passes for "group work" in classrooms is Active at best (students dividing tasks and working in parallel) rather than Interactive (Training-derived).

Fiorella (2023) proposed "sense-making" as the common mechanism underlying all generative learning strategies — summarizing, mapping, drawing, self-explaining, teaching, enacting (Abstract-verified). This unifies the ICAP hierarchy with CLT's emphasis on schema construction: the cognitive work that makes Constructive and Interactive modes effective is the same schema- building work that worked examples support by different means. Sailer et al. (2024) have begun investigating additional boundary conditions, including domain and task-type moderators (Abstract-verified).

The ICAP framework's practical power for this review lies in its independence from the direct-instruction-vs-inquiry framing. A lecture becomes more effective when students self-explain (Constructive) rather than passively listen (Passive) — this is true regardless of whether the instruction is "direct" or "inquiry-based." An inquiry activity becomes less effective when students divide tasks mechanically (Active) rather than genuinely co-construct understanding (Interactive). The critical variable is not who initiates the activity or what the activity is called, but what cognitive processing the activity demands of the learner.

6.4 ENGELMANN'S DIRECT INSTRUCTION: EVIDENCE AND LIMITS

Siegfried Engelmann's Direct Instruction (capital letters distinguish his specific program from generic direct instruction) is the most evidence-supported and least adopted instructional program in education. Reading Engelmann and Carnine's *Theory of Instruction* (1991) at primary-text depth reveals a rigorous logical framework. The core idea is "faultless communication" — designing instructional sequences that are "analytically or logically capable of transmitting the concept or skill to any learner who possesses certain minimal attributes" (§0.3.6)•. The design unit is the *communication*, not the learner. "If the learner hasn't learned, the teacher hasn't taught" — not as rhetoric but as diagnostic methodology: "the value of the initial hypothesis of the problem (that the teaching is the sole cause of the learner's problem) is that it requires us to rule out the possibility that instructional variables could account for learner failure" (§0.5.4)•.

The examples-and-non-examples design principle is genuinely powerful: "If we design a negative example so that it is highly similar to a positive example, we rule out the greatest number of possible interpretations. Only the difference between the positive and the negative can account for one example being positive and the other negative" (§0.5.26)•. This is a logically sound principle for concept teaching that has been validated by the worked-example research.

The evidence base is unusually strong. Stockard et al. (2018) meta-analyzed 328 studies spanning 50 years and found that Engelmann's DI produced positive outcomes across all student populations and all outcome measures except affective outcomes (Abstract-verified). Project Follow Through — the largest educational experiment in U.S. history — found DI produced the best results not just in basic skills but in problem-solving and self-esteem (Training-derived).

Yet Engelmann's DI has critical limitations that the primary text makes visible. First, the theory has *no treatment* of ill-structured domains. The framework assumes all target content can be logically analyzed into unambiguous positive/negative example sets. Engelmann acknowledges that some concept boundaries are vague — "when a shoe becomes a not-shoe is not known" (§0.5.24)• — but treats this as a special case, not as a fundamental challenge to the framework. Writing, design, ethical reasoning, and artistic practice do not fit the examples-and-non-examples paradigm.

Second, the dismissal of discovery learning is absolute: "A logical analysis of any discovery situation reveals why it is far inferior to a more structured format" (§0.21.57)●. This does not engage the productive-failure evidence — the finding that structured struggle *before* instruction produces deeper learning than instruction alone. Engelmann concedes only that "the learner becomes practiced in discovery only through discovering" (§0.21.58)●, but immediately reframes this as a design problem within the DI paradigm.

Third, the compliance cascade. Kohn (1999) argued that "top-down control of schools by legislators and other policy makers" leads to "top-down control of classrooms by teachers" which denies students "the chance to direct their own learning" (§8.1.10)●. This is not a replication failure but an implementation pathology that the experimental literature does not measure. Stockard et al.'s affective-outcomes exception — the one area where DI did not produce positive results — is consistent with Kohn's critique. Deci, Koestner, and Ryan (1999) established that tangible rewards undermine intrinsic motivation (Training-derived); the controlling features of scripted DI programs may produce analogous effects at the institutional level.

The honest assessment: Engelmann's DI is probably the most effective approach available for rapidly building foundational knowledge and skills in well-structured domains. Its evidence base — 328 studies, 50 years, consistently positive effects across populations — should not be ignored by anyone designing instruction for novice learners in content that can be decomposed into clear concepts and procedures. But it is not a complete instructional model, and its scope is limited to content that can be logically decomposed into unambiguous example sets. The compliance cascade is a genuine concern, not an aesthetic objection — the affective-outcomes exception in Stockard's own meta-analysis is the evidence-internal signal that something important is missing from the DI picture. And DI's undeniable success on achievement tests does not address the question that W2-008's convergence map raises: whether it develops the learners that a curriculum ultimately needs — autonomous, self-regulating individuals capable of judgment in novel situations, embedded in warm reciprocal relationships. The staged model places DI's contributions at Stages 1–2 and draws on other traditions for Stages 3–5 precisely because DI itself provides no guidance for developing judgment, metacognition, or character.

6.5 THE THREE-TRADITIONS MAP

Reading the primary texts reveals that instructional design research comprises not two traditions (as the binary debate suggests) but at least three:

1. **Cognitive-psychology tradition** (CLT, worked examples, expertise reversal, Sweller/Kalyuga/van Merriënboer) — analyzes instruction in terms of cognitive load and schema construction. 2. **Learning-sciences tradition** (inquiry, PBL, productive failure, Hmelo-Silver/Kapur/Chi) — analyzes instruction in terms of knowledge construction, collaboration, and metacognition. 3. **Instructional-practice tradition** (Rosenshine, Engelmann, Lemov, Sherrington) — analyzes instruction in terms of observable teacher behaviors and student outcomes.

These traditions rarely cite each other, and the non-communication is not just an oversight — it reflects genuinely different research methodologies, publication venues, and intellectual values. Sweller et al. (2019) do not mention Kapur; their 20-year retrospective engages the cognitive-psychology tradition exhaustively but the learning-sciences tradition barely at all. Kapur (2024) does not mention Sweller by name; his productive-failure work frames itself against "direct instruction" as a monolith without engaging CLT's specific mechanisms. Rosenshine (2012) cites cognitive science (working memory, schema theory) but not PBL or inquiry learning. Engelmann and Carnine (1991) cite none of the above — their theoretical framework is entirely self-contained,

grounded in logical analysis of instructional communication rather than cognitive-psychological theory. Lemov (2021) cites Rosenshine and Willingham but not Sweller, Kapur, or Chi.

The binary debate is partly an artifact of this non-communication: each tradition talks to itself and characterizes the others through caricature. The cognitive- psychology tradition sees inquiry as "unguided"; the learning- sciences tradition sees direct instruction as "passive transmission"; the practitioner tradition sees both academic camps as disconnected from classroom reality. The staged- instruction model in Section 7 is an attempt to integrate all three — drawing on CLT's explanation of *why* explicit instruction works for novices, the learning sciences' explanation of *why* exploration enhances conceptual understanding, and the practitioner tradition's knowledge of *how* the transitions actually work in classrooms.

7 THE STAGED-INSTRUCTION MODEL

This is the central synthetic contribution of this review. The binary framing — direct instruction vs. inquiry — is misleading. The evidence supports an expertise-adaptive sequence with at least five distinguishable stages. Each stage is defined by the learner's current knowledge state, the dominant instructional mode, and the transition signal that indicates readiness for the next stage. The model draws on all three traditions identified in Section 6: cognitive psychology provides the rationale for Stages 1–2, the learning sciences provide the rationale for Stages 3–4, and the practitioner tradition provides the operational detail for transitions.

Three important caveats before the stage-by-stage description. First, the stages describe a general arc, not a rigid sequence. As Lemov notes, "your students may be fairly knowledgeable; then they start a new unit on Monday and move back to square one again" (2021, §11.5.2)●. A learner may be at Stage 4 in one topic and Stage 1 in another within the same course. Second, the duration of each stage varies enormously by domain, content difficulty, and learner background — no timeline is implied. Third, the model is grounded in the evidence reviewed in Sections 2–6, but the integration itself — the claim that these stages form a coherent sequence — is this review's synthetic contribution, not a position endorsed by any of the cited researchers. Sweller, Kapur, and Engelmann would each object to aspects of the integration.

7.1 STAGE 1: EXPLICIT INSTRUCTION WITH WORKED EXAMPLES

Learner state. Genuine novice. No relevant schemas in long-term memory. Working memory is the only resource.

What instruction looks like. Present new material in small steps with practice after each step (Rosenshine, 2012, Principle 2)●. Provide worked examples with step-by-step explanations — the worked-example effect is one of CLT's most robust findings for this population. Use Engelmann's examples/non-examples for concept teaching: "If we design a negative example so that it is highly similar to a positive example, we rule out the greatest number of possible interpretations" (Engelmann & Carnine, 1991, §0.5.26)●. Check for understanding continuously — Sherrington calls this "the single biggest common area for improvement" in observed teaching (2019, Strand 2)●. Obtain a high success rate (80%) to prevent error consolidation (Rosenshine, 2012, p. 17)●. The model-then- test sequence: teacher demonstrates first five examples, then tests the next six (Engelmann & Carnine, 1991, §0.5.3)●.

Evidence base. KSC 2006/CKS 2012 (Verified direct for CKS); Rosenshine 2012 (Verified direct); Engelmann & Carnine 1991 (Verified direct); van Merriënboer & Kirschner 2018 completion strategy (Verified direct); Stockard et al. 2018 (Abstract-verified). Tetzlaff et al. 2025 meta-analysis: low- prior-knowledge learners benefit from high-assistance instruction at d = 0.505 (Abstract-verified).

Transition signal OUT of Stage 1. Unsupported performance exceeds 80% on task-class assessments (Rosenshine; van Merriënboer's horizontal standards matrix). Students can solve standard problems without referring to worked examples — the completion strategy has been fully faded (van Merriënboer, §0.10.47)●. Teacher judgment: "knowing the material, knowing how to break it down, and knowing the students" (Sherrington, 2019, Strand 4)●.

Scope condition. This stage applies most directly to well-structured content. For ill-structured domains, Stage 1 is compressed or modified — see Section 8.

7.2 STAGE 2: SCAFFOLDED PRACTICE WITH FADING

Learner state. Developing schemas. Can solve standard problems but schemas are fragile and not yet automated. Beginning to recognize problem types.

What instruction looks like. Completion problems: partially worked examples where students fill in increasing numbers of steps (van Merriënboer, §0.8.34)●. The 4C/ID scaffolding-fading sequence within task classes. Directive questioning that teaches rigorous thinking — Lemov's "Stretch It" treats "correct answers as a step in the learning process" rather than endpoints (Lemov, 2021, §10.15.31)●. Part-task practice for recurrent skills needing automation (4C/ID component 4). Supportive information — domain models and cognitive strategies — for non-recurrent skills (4C/ID component 2).

Evidence base. Van Merriënboer & Kirschner 2018 (Verified direct); Renkl & Atkinson 2003 fading-guidance strategy (Training-derived); Sweller et al. 2019 guidance-fading effect (Verified direct); Lemov 2021 directive-to-nondirective progression (Verified direct).

Transition signal OUT of Stage 2. The expertise reversal threshold: additional worked examples or scaffolding become redundant and interfere with learning (Sweller et al., 2019, p. 277)●. Students can articulate the problem-solving process — not just execute it — answering "process questions" (Rosenshine, 2012, p. 14)●. Students possess sufficient prior knowledge to generate suboptimal but relevant solutions to novel problems — Kapur's accessibility criterion (§14.2.1)●.

The transition signal is the key unsolved problem. No reliable classroom-level diagnostic exists. Sherrington acknowledges this is "a subtle skill, a central element of teacher expertise" (2019, Strand 4)●. The 80% threshold is a rough heuristic, not an operational criterion.

7.3 STAGE 3: PRODUCTIVE FAILURE / STRUCTURED EXPLORATION

Learner state. Has enough schemas for productive activation. Can generate relevant (if suboptimal) approaches to novel problems. Working memory is augmented by organized long-term memory.

What instruction looks like. The PS-I sequence: problem- solving phase BEFORE instruction phase (Kapur, 2024, §7.5.3)●. Tasks are "challenging but accessible" — no technical jargon, invite multiple solution representations (§14.2.1, §14.4.1)●. Students generate and compare multiple approaches; the generation itself is the learning mechanism. The instruction phase follows: teacher assembles correct concepts, explicitly connecting to students' generated solutions (Kapur's Assembly principle, §8.0.6)●. Culture of Error: mistakes are "a first, positive, and often critical step toward getting it right" (Lemov, 2021, §9.10.4)●.

For writing instruction: Hillocks' environmental mode — structured problem-solving activities with clear objectives and peer interaction — achieves ES = .44, versus presentational mode at .02 and natural process at .19 (Hillocks, 1986, §0.8.210)●.

For design education: Brousseau's *adidactical situations* — "the student knows very well that the problem was chosen to help her acquire a new piece of knowledge, but she must also know that this knowledge is entirely justified by the internal logic of the situation and that she can construct it without appealing to didactical reasoning" (Brousseau, 1997, §1.0.7)●.

Evidence base. Kapur 2024 meta-analysis: up to 3x effect size over DI for conceptual understanding (Verified direct). Hillocks 1986: environmental mode ES = .44 (Verified direct). Brousseau

1997: milieu as designed environment for productive cognitive conflict (Verified direct). Schwartz & Bransford 1998: preparation for future learning (Training-derived).

Transition signal OUT of Stage 3. Students can independently frame problems — identifying what kind of problem they face without being told (Chi, Feltovich & Glaser, 1981: deep-structure recognition, as discussed in W2-009 §II). They ask their own questions and generate their own variations. They can articulate not just what they did but *why* particular approaches work and others do not.

Connection to Brousseau. Stage 3 is the beginning of *devolution* — the teacher is transferring responsibility for the problem to the student, not just fading scaffolding.

7.4 STAGE 4: GUIDED INQUIRY / DESIGN STUDIO

Learner state. Deep schemas, some automated sub-skills. Can frame problems, not just solve them. Recognizes deep structure across problem types. Beginning to develop judgment (Layer 3 of the competence stack).

What instruction looks like. Open-ended problems with multiple acceptable solutions and genuine uncertainty. Teacher role shifts from instructor to coach. Schön's (1987) "joint experimentation" — coach and student collaboratively explore a problem space: "Dani suggests many ways — not one best way" (§0.20.104)●. Nondirective questioning replaces directive questioning: Lemov describes "working in more nondirective prompts over time" (2021, §10.24.10)●. Cognitive strategies and domain models become primary learning supports (4C/ID supportive information). Schön's "virtual worlds" — practice environments that preserve cognitive demands but reduce irreversibility (1983, §1.18.70–75)●.

Evidence base. Schön 1983, 1987 (Verified direct for both); van Merriënboer & Kirschner 2018 task classes at higher complexity levels (Verified direct); Brousseau 1997 adidactical situations (Verified direct); Chi ICAP: Interactive

Constructive modes (Abstract-verified).

Transition signal OUT of Stage 4. The student can set problems, not just solve them — "problem setting, the process by which we define the decision to be made, the ends to be achieved, the means which may be chosen" (Schön, 1983, §1.4.38)●. The student has developed *repertoire* — a store of exemplars guiding recognition and action without explicit deliberation. The student can reflect-in-action: "thinking what they are doing and, in the process, evolving their way of doing it" (Schön, 1983, §1.4.54)●.

7.5 STAGE 5: INDEPENDENT PRACTICE / REFLECTIVE EXPERTISE

Learner state. Expert or near-expert. Extensive schemas, automated procedures, developed judgment. Capable of reflection-in-action. Can construct "a new theory of the unique case" (Schön, 1983, §1.4.66)●.

What instruction looks like. Self-directed learning with periodic expert consultation. Second-order scaffolding: shared control over task selection (van Merriënboer, §0.21.1)●. Deliberate practice targeting specific weaknesses (Ericsson)○. The teacher is available but not directing — the student drives the learning process. For professional education: residency, studio practice, clinical supervision.

Evidence base. Ericsson 2004 deliberate practice (Training-derived); van Merriënboer & Kirschner 2018 second-order scaffolding (Verified direct); Schön 1983, 1987 reflective practitioner (Verified direct).

7.6 A WORKED EXAMPLE: STATISTICS (WELL-STRUCTURED DOMAIN)

Consider teaching the concept of standard deviation to secondary-school students.

Stage 1 (1–2 lessons): Teacher presents worked examples of how to calculate mean deviation, with step-by-step annotation. Students work through completion problems — first three steps given, then two, then one. Engelmann's examples/non-examples: data sets that are/aren't "spread out" to establish the concept boundary.

Stage 2 (2–3 lessons): Students solve varied problems with decreasing scaffolding. Process questions: "Why did you subtract from the mean?" Part-task practice on calculation fluency.

Stage 3 (1–2 lessons): Productive failure task — students are given data on two basketball players and asked to determine "who is more consistent." They generate multiple measures (range, mean deviation, visual comparisons) and fail to find the canonical answer. Instruction follows: teacher assembles students' approaches, shows what each captures and misses, builds toward standard deviation. This is Kapur's paradigm case, with effect sizes up to 3x compared to I-PS.

Stage 4 (2–3 lessons): Students design their own statistical investigations — choosing data sets, selecting appropriate analyses, interpreting results. Teacher coaches: "What question are you trying to answer? Does this measure capture what you care about?"

Stage 5 (ongoing): Independent statistical projects. Students identify questions, gather data, choose methods, interpret results, present findings. Teacher available for consultation.

7.7 A WORKED EXAMPLE: ESSAY WRITING (ILL-STRUCTURED DOMAIN)

Consider teaching persuasive essay writing to secondary-school students.

Stage 1 — modified: Direct instruction of essay-writing rules has near-zero effect (Hillocks: presentational mode ES = .02)●. Instead, Stage 1 in this domain looks like Schön's "Follow me!" — holistic demonstration and imitation, not decomposed worked examples. The teacher models the writing process through think-aloud, making the reasoning behind choices visible: "I'm choosing this example because my reader is likely to object that..." Students write parallel texts following the demonstrated pattern. Model texts (exemplars) serve the role that worked examples serve in mathematics — but with multiple "correct" solutions.

Stage 2: Students analyze exemplar texts with guided questions: "What is the author's claim? What evidence supports it? Why did they address the counterargument here?" This is the ill-structured equivalent of completion problems — partial analysis that students extend.

Stage 3 — primary mode: Hillocks' environmental mode — structured problem-solving activities with clear criteria and peer interaction. Students are given a claim and asked to construct the strongest possible argument, then the strongest possible counterargument, then integrate both. "Principles are not simply announced and illustrated [...] rather, they are approached through concrete materials and problems" (Hillocks, 1986, §0.8.135)●. Peer response groups provide structured feedback. This stage may consume the majority of instructional time in writing — the environmental mode is the dominant mode, not a transitional one.

Stage 4: Students choose their own topics and design their own arguments. Teacher-as-coach provides targeted feedback through writing conferences — Schön's joint experimentation applied to text. Nondirective: "What effect are you trying to produce here? Is this achieving it?"

Stage 5: Independent writing projects with editorial feedback. The student drives topic selection, argument design, revision process. Teacher available as reader and critic.

7.8 WHAT THE EVIDENCE DOES NOT SUPPORT

The staged model rests on genuine evidence, but several claims must be flagged:

1. **No specific timeline** for each stage. Duration varies by domain, content difficulty, and learner background. 2. **No single transition threshold.** The 80% heuristic is from one fourth-grade math study; it is not generalizable. 3. **PF is not for genuine novices.** Kapur's own design principle requires accessible tasks — students must have enough prior knowledge for productive generation. 4. **The stages are not strictly sequential within a course.** A teacher may use Stage 1 for new content and Stage 3 for previously learned content simultaneously. "Your students may be fairly knowledgeable; then they start a new unit on Monday and move back to square one again" (Lemov quoting Sweller, 2021, §11.5.2)●. 5. **The transition signals for Stages 3→4 and 4→5 rest on thinner evidence** than those for 1→2 and 2→3. Deep- structure recognition (Chi et al., 1981) and problem-setting (Schön, 1983) are theoretically grounded but not operationalized as classroom diagnostics.

7.9 THE STAGED MODEL AND THE COMPETENCE STACK

Mapping the stages against Applied Pedagogy's five-layer competence stack (knowledge, skill, judgment, metacognition, character/disposition) reveals an important asymmetry:

- **Stages 1–2** primarily develop Layers 1–2: knowledge and skill. Worked examples build domain knowledge; scaffolded practice develops skill. - **Stage 3** bridges Layers 2–3: productive failure develops conceptual depth that supports the transition from skill to judgment. The Awareness mechanism develops metacognition (Layer 4). - **Stage 4** develops Layer 3: judgment through guided inquiry in authentic, ambiguous situations. The design studio, the environmental mode, the adidactical situation — all place learners in contexts requiring judgment. - **Stage 5** develops Layers 4–5: metacognitive self-regulation and the epistemic dispositions (intellectual honesty, tolerance for uncertainty) that characterize expert practice.

W2-009 described the Layer 2→3 transition as requiring "accumulated experience with varied, consequential situations in environments that provide valid feedback." The staged model specifies what this looks like instructionally: the transition from scaffolded practice (Stage 2) through productive failure (Stage 3) to guided inquiry (Stage 4) is the designed sequence of experiences that builds the representations supporting judgment. The design is not "teach judgment directly" but "create the conditions under which judgment develops."

8

INSTRUCTIONAL DESIGN IN ILL-STRUCTURED DOMAINS

This section addresses v1 Gap 1 — the field's most significant gap for curriculum design. v1 flagged the near-total absence of guidance for instructional design in domains where problems have unknown elements, multiple acceptable solutions, and criteria requiring judgment. v2 engages the domain-specific traditions directly.

8.1 THE CONVERGENCE OF INDEPENDENT TRADITIONS

Five ill-structured domain traditions were assessed in Session 2's scorecard: writing pedagogy (rooted in American composition research), design education (rooted in American and British professional education), French *didactique* (rooted in French mathematics education), clinical reasoning training (rooted in Canadian and Dutch medical education), and artistic apprenticeship (rooted in the European conservatory tradition). These traditions developed independently — in different countries, different decades, different languages, and from different theoretical frameworks. They have different publication venues, different intellectual heroes, and different methodological standards. Yet when their findings are compared, they converge on three discoveries that are all the more striking for having been made independently:

Finding 1: Direct instruction of rules fails in ill-structured domains. Hillocks (1986): presentational mode (direct teaching of writing rules) produces an effect size of .02 (§0.8.210)●. Schön (1987): "a designlike practice is learnable but is not teachable by classroom methods" (§0.20.115)●. Brousseau (1997): the Topaze effect — scaffolding that simplifies until the learning target disappears entirely (§1.0.2)●. Engelmann's own framework acknowledges the limit: "when a shoe becomes a not-shoe is not known" (§0.5.24)●. Three independent traditions, three independent discoveries that telling does not work when there is nothing unambiguous to tell.

Finding 2: Unstructured exploration also fails. Hillocks (1986): the "natural process" mode (free writing with minimal structure) produces ES = .19 — better than presentational but far below the environmental mode (§0.8.210)●. "Free writing [...] may reinforce the 'what next' strategy" (§0.8.242)●. Schön (1987): the "mystery and mastery" dynamic where unsupported students are lost in the complexity of practice. KSC (2006): novices learn "almost nothing" from unguided search (Training-derived). The answer is not "let them discover" any more than it is "tell them the rules."

Finding 3: Structured activity with clear objectives and coaching works. Hillocks' environmental mode: structured problem-solving with peer interaction and specific objectives achieves ES = .44 — more than four times the presentational mode and three times the natural process mode (§0.8.257)●. Schön's joint experimentation: coach and student collaboratively explore a problem space (1987, §0.20.180)●. Brousseau's adidactical situations: designed environments that generate productive cognitive conflict (1997, §1.0.7)●. Kapur's productive failure: designed tasks that generate productive struggle followed by instruction (2024, §7.5.3)●. All involve carefully designed tasks that generate productive struggle, followed by expert guidance.

8.2 THE ENVIRONMENTAL MODE AS DESIGN PRINCIPLE

Hillocks' environmental mode deserves extended treatment because it is the strongest evidence-based approach to ill-structured instruction. Four assumptions underlie it: "(1) teaching can and should actively seek to develop identifiable skills in learners; (2) these skills are developed by using them orally before using them in writing; (3) one major function of prewriting activity is to develop those skills; (4) the use of such skills [...] is often complex, and therefore may require collaboration with and feedback from others" (Hillocks, 1986, §0.8.138)•.

The key insight is the declarative/procedural distinction as it applies to ill-structured content: "Traditional approaches to teaching composition have concentrated on declarative knowledge of grammar [...]. Research examined in this review indicates clearly that approaches which focus on procedural knowledge (e.g., sentence combining, scales, inquiry) are more successful than those which focus on declarative knowledge" (§0.8.243)•. In well-structured domains, declarative knowledge (rules, formulas, principles) is often the starting point for instruction because it can be directly applied. In ill-structured domains, declarative knowledge about the domain (grammar rules, design principles, ethical frameworks) has near-zero transfer to practice. The environmental mode develops procedural knowledge through structured activity — not by announcing principles but by engaging learners in tasks that embody them.

8.3 DEVOLUTION: THE ILL-STRUCTURED EQUIVALENT OF SCAFFOLDING-FADING

In well-structured domains, the primary instructional transition is scaffolding-fading: reducing support from a fixed problem. In ill-structured domains, Brousseau's concept of *devolution* provides the better model: the teacher transfers *responsibility for the problem itself* to the student. "Devolution is the act by which the teacher makes the student accept the responsibility for an (adidactical) learning situation or for a problem, and accepts the consequences of this transfer of this responsibility" (Brousseau, 1997, §5.1.4)•.

The difference is not merely terminological. Scaffolding-fading assumes the problem remains constant while support decreases. Devolution changes *what the student is responsible for*. At early stages, the teacher structures the milieu — designs the situation, selects the problem, defines the constraints. At later stages, the student accepts responsibility for the situation itself — defining the problem, choosing the approach, evaluating the outcome. Schön's transition from "Follow me!" (novice phase — the student imitates the master) to "joint experimentation" (advanced phase — coach and student collaboratively explore) is the same structural transition described in different terms (Schön, 1987, §0.20.178–180)•.

8.4 DESIGN EDUCATION AND THE STUDIO MODEL

Schön's analysis of the architectural design studio (1983, 1987) provides the most detailed protocol for instruction in ill-structured domains. The Quist/Petra case (1983, §1.18.6–7)• — a master architect coaching a student through a design problem — illustrates the core mechanism: "reflective conversation with the situation." The master reframes the student's problem, demonstrates design moves, and listens to the situation's "back-talk." Each move is "a local experiment which contributes to the global experiment of reframing the problem" (§1.18.7)•.

The studio model's pedagogical structure maps onto the staged model with modifications. Stage 1 looks like "Follow me!" — holistic demonstration where the student "cannot at first understand what he needs to learn" (Schön, 1987, §0.20.44)•. The "learning predicament" is real: the student

must begin before they can understand the goal. Stage 4 looks like "joint experimentation" — the transition signal is that the student can "say what effects she would like to produce" (Schön, 1987, §0.20.180)•, at which point the coach shifts from demonstration to collaborative exploration.

The "virtual world" concept is particularly valuable for instructional design: practice environments that preserve the cognitive demands of real practice while removing the irreversibility of real consequences. "No move is irreversible. The designer can try, look, and by shifting to another sheet of paper, try again" (Schön, 1983, §1.18.71)•. This connects to van Merriënboer's emphasis on psychological fidelity in simulation design — and provides a design principle for ill-structured practice across domains: create contexts where learners can take consequential actions without irreversible consequences.

8.5 CLINICAL REASONING AND SCRIPT THEORY

The medical education tradition provides the clearest evidence for instructional design at the skill-to-judgment transition — the Layer 2→3 shift in the competence stack. Schmidt and Boshuizen's (1993) illness-script theory describes how biomedical knowledge becomes "encapsulated" into clinical knowledge as medical students gain experience — a specific instance of the surface-to-deep-structure reorganization that Chi et al. (1981) documented in physics (Training-derived). The novice medical student applies biomedical knowledge step by step (surface features: "the patient has fever and elevated white count → infection"). The experienced clinician recognizes patterns instantly ("this looks like appendicitis") without explicitly rehearsing the biomedical reasoning. The knowledge has not been lost but encapsulated — compressed into illness scripts that function as schemas for clinical judgment.

This encapsulation process is instructionally significant because it describes the mechanism by which explicit knowledge transforms into judgment. It cannot be directly taught — no amount of lecturing about illness scripts produces the encapsulation that experience produces. But it can be designed for: case-based instruction with progressive complexity creates the conditions under which encapsulation occurs. The early cases are simple (single-system, classic presentation); the later cases are ambiguous (multi-system, atypical presentation). The progression is the 4C/ID task-class architecture, refined by a century of medical education.

Lubarsky et al.'s (2015) work on script-concordance testing (SCT) provides an assessment instrument for judgment under uncertainty — testing whether a learner's judgment patterns converge with expert panels (Abstract-verified). SCT does not test whether the learner *knows* the right answer (there may be no single right answer) but whether the learner's judgment under uncertainty *moves in the same direction* as expert judgment. This is an assessment tool designed specifically for the ill-structured, judgment-laden performance that standardized tests cannot measure — and it provides a possible operationalization of the Stage 3→4 transition signal (when learner judgment begins to converge with expert judgment, the learner is ready for less-structured practice).

Medical education has engaged CLT and 4C/ID at implementation scale in a way K-12 has not. The case-based instruction that is standard in medical schools is, functionally, the 4C/ID whole-task approach with task classes of increasing complexity. The OSCE (Objective Structured Clinical Examination) is a form of authentic assessment designed for ill-structured performance. The progression from basic science → clinical clerkships → residency → independent practice is a staged- instruction model refined over a century of institutional experience.

8.6 THE FRENCH *DIDACTIQUE* CONTRIBUTION

Brousseau's theory of didactical situations (1997) and Chevallard's didactic transposition (1985/1991) provide conceptual tools the Anglo-American tradition lacks.

The *fundamental paradox* of instruction: "everything that [the teacher] undertakes in order to make the student produce the behaviours that she expects tends to deprive this student of the necessary conditions for [...] learning" (Brousseau, 1997, §1.2.2)•. This is not a bug to be fixed but a structural feature of instruction. Every act of teaching risks eliminating the learning it aims to produce.

The *Topaze effect* and *Jourdain effect* are instructional pathologies that CLT cannot diagnose. The Topaze effect occurs when "by choosing easier and easier questions, the teacher tries to achieve the optimum meaning for the maximum number of students. If the target knowledge disappears completely, we have the Topaze effect" (§1.0.2)•. The Jourdain effect occurs when "the teacher agrees to recognize the indication of an item of scientific knowledge in the student's behaviour or answers, even though these are in fact motivated by ordinary causes and meanings" (§1.0.3)•. The Topaze effect is the failure mode of scaffolding (reducing difficulty until nothing is learned); the Jourdain effect is the failure mode of inquiry (over-interpreting student responses as understanding). Both are endemic to ill-structured domains where the criteria for "correct" performance are difficult to specify.

Epistemological obstacles are a concept absent from CLT: "An obstacle is a piece of knowledge or a conception, not a difficulty or a lack of knowledge. This piece of knowledge produces responses which are appropriate within a particular, frequently experienced, context. But it generates false responses outside this context" (Brousseau, 1997, §2.0.22, citing Duroux)•. Knowledge as barrier, not absence — the five-paragraph essay is an epistemological obstacle for the student who must learn to write for professional audiences; the cookbook lab procedure is an epistemological obstacle for the student who must learn to design experiments.

Chevallard's *didactic transposition* (1985/1991) adds a structural analysis: taught knowledge necessarily differs from scholarly knowledge due to the institutional demands of schooling — decontextualization, sequencing, assessment requirements. "Knowledge is subject to a set of transformations adaptives qui vont le rendre apte à prendre sa place parmi les objets d'enseignement" (Chevallard, §0.9.6, original French)•. This analysis explains why well-structured school tasks (textbook problems, five-paragraph essays) often fail to develop the competence that ill-structured professional practice demands: the transposition has stripped the ambiguity, judgment, and context-sensitivity that make the knowledge meaningful.

8.7 NON-ENGLISH TRADITIONS: GALPERIN, KLAFKI, LESSON STUDY

Galperin's stepwise formation of mental actions (via Engeness, 2020)•provides a six-phase model: motivation → orientation → materialised action → communicated thinking → dialogical thinking → acting mentally. The model was developed in the 1950s–60s within Soviet activity theory, entirely independently of CLT, yet the structural parallels are remarkable. The progression from materialised action (manipulating physical objects or written procedures) through communicated thinking (articulating the process aloud to others) to dialogical thinking (internalizing through internal dialogue) and finally acting mentally (automatized performance) describes the same externalization-to-internalization trajectory that CLT's worked-example-to-independent-practice describes — but with explicit social-speech phases that CLT does not theorize. The communicated-thinking phase — where the learner articulates the process to another person — corresponds to

the ICAP framework's Constructive or Interactive modes: the cognitive work of articulation is itself a learning mechanism.

Galperin's three types of orientation are particularly illuminating for the staged model. Type 1 orientation is incomplete and trial-and-error — the learner proceeds without a clear model of the goal or the process. This corresponds to unguided discovery, and Galperin (like Kirschner et al.) considered it inefficient. Type 2 orientation is complete and teacher-provided — the learner receives a detailed orientation basis (a schema, a procedure, a worked example) before beginning. This corresponds to explicit instruction with worked examples. Type 3 orientation is complete but *learner-constructed* — the learner builds the orientation basis through guided activity. This corresponds to productive failure and guided inquiry. Galperin argued that Type 3 produces the most transferable knowledge — the learner who constructs the orientation basis understands it more deeply than the learner who receives it. This is, independently, the productive-failure argument: the generation of (even incorrect) orientation precedes and enhances the reception of canonical knowledge.

The design principle that "feedback should be faded" (Engeness DP6) parallels the expertise reversal effect — yet another independent discovery of the same instructional principle from a different theoretical tradition and a different continent.

Klafki's categorical *Bildung* (via Sjöström & Eilks, 2020)• provides a pre-design step absent from Anglo-American ID: the *Didaktik* analysis asks five questions before instruction is designed — exemplarity (is this content representative?), present significance (does it matter to the student now?), future significance (will it matter later?), content structure (how is it organized?), and accessibility (how can the student encounter it?). This is a content- selection framework that addresses the "what" before the "how" — and connects to W2-008's normative question about what instruction is *for*. As Sjöström and Eilks note, *Didaktik* focuses on aims while Anglo-American ID focuses on methods; the two traditions need each other.

Japanese lesson study (Takahashi & McDougal, 2016)• provides a *process* for arriving at instructional designs through iterative collaborative refinement. The six characteristics of Collaborative Lesson Research — clear research purpose, *kyouzai kenkyuu* (study of instructional materials), written research proposal, live research lesson, post-lesson discussion, sharing of results — describe a professional-learning methodology compatible with any theoretical framework. Lesson study is not a competing model of instruction but a process for developing and refining instruction within CLT, 4C/ID, PF, or any other lens.

8.8 WHAT THE ILL-STRUCTURED GAP CLOSING ACHIEVED

v1 Gap 1 flagged the near-total absence of guidance for ill-structured domains. v2 narrows the gap substantially but does not close it. What was achieved:

- **The convergence finding** — three independent traditions (writing pedagogy, design education, French *didactique*) independently discovered that direct instruction of rules fails and structured activity with coaching works in ill-structured domains. - **The environmental mode** as a concrete, evidence-based alternative to both direct instruction and unguided exploration for ill-structured content. - **Devolution** as the ill-structured equivalent of scaffolding-fading. - **The Topaze/Jourdain effects** as diagnostic tools for instructional pathologies. - **The staged model's ill-structured variant** with modified stages and domain-specific worked examples.

What remains open:

No cross-domain meta-analysis of ill-structured instruction exists. The convergence finding — that structured activity with coaching works across writing, design, and clinical reasoning — is an

inference from domain-specific literatures, not a meta-analytic result. Whether the environmental mode transfers from writing to design to clinical reasoning with comparable effect sizes is genuinely unknown. The domains differ in important ways: writing produces a tangible artifact that can be revised; design involves visual-spatial reasoning; clinical reasoning operates under time pressure and asymmetric consequences. These differences may require domain-specific adaptations that resist the unification the convergence finding suggests.

Exemplar-based instruction is under-theorized. In ill-structured domains, "worked examples" become exemplars — model texts in writing, design precedents in architecture, teaching cases in medicine, performance recordings in music. The CLT worked-example literature provides precise guidance on how examples should be designed for well-structured domains (step-by-step annotation, fading, interleaving). No comparable literature exists for exemplars in ill-structured domains. How should exemplars be selected — should they represent excellent performance, typical performance, or deliberately flawed performance? How should they be presented — with expert commentary, with guided analysis questions, or with peer discussion? Should learners imitate exemplars (the "Follow me!" phase) or diverge from them (the creative development phase)? The writing-pedagogy community uses "mentor texts" as instructional anchors; the design community studies "precedents"; the medical community analyzes "teaching cases." Whether these are the same instructional tool with different names, or fundamentally different tools requiring different design principles, is a question the field has not addressed.

The Stage 2→3 transition signal in ill-structured domains is different from well-structured domains and not well operationalized. In mathematics, the transition is marked by the ability to solve standard problems without scaffolding (80% accuracy). In writing, what counts as a "standard problem" and what counts as "accuracy" are themselves ambiguous. When should instruction shift from guided exemplar analysis to environmental-mode tasks?

The Topaze and Jourdain effects need operationalization. They are powerful diagnostic concepts for instructional pathologies — Brousseau identified them from mathematics classroom observation in the 1980s. But practitioners in other domains have no reliable way to detect them in real time. Could classroom observation instruments be developed that enable teachers (or AI-powered classroom analytics) to recognize when scaffolding has eliminated the learning target (Topaze) or when student responses are being over-interpreted as understanding (Jourdain)?

9

THE AUTONOMY-STRUCTURE TENSION

v1 Gap 8 flagged the tension between CLT's prescription for explicit instruction and SDT's prescription for autonomy support. v2 treats this as a design constraint, not a philosophical debate.

9.1 THE TENSION IS NOT CONSTANT

The autonomy-structure tension changes character at each stage of the model:

At **Stage 1**, structure dominates by cognitive necessity — novices need guidance because their working memory cannot simultaneously search for solutions and learn from the process. But the risk is the compliance cascade: "top-down control of schools by legislators and other policy makers" leads to "top-down control of classrooms by teachers" which denies students "the chance to direct their own learning" (Kohn, 1999, §8.1.10)●. Stage 1 instruction can be delivered in a controlling way (this is Kohn's target) or in an autonomy-supportive way. The difference lies not in the *content* of instruction — which CLT constrains — but in the *delivery*. Autonomy-supportive delivery at Stage 1 means: providing rationale for why this content matters ("we're learning standard deviation because you'll need it to evaluate whether differences in data are meaningful"), offering choice within structure ("you can choose which data set to analyze"), acknowledging difficulty ("this is genuinely hard — it's normal to struggle with this"), and using invitational language ("let's try this approach" rather than "do this"). Ahmadi et al. (2023) catalogued 57 specific autonomy-supportive behaviors from SDT intervention research (Abstract-verified); which of these are compatible with the cognitive demands of high-quality Stage 1 instruction has not been investigated.

At **Stage 2**, fading scaffolding incrementally increases autonomy. The transition IS the resolution — as scaffolding fades, the learner takes on more cognitive work and more control over the learning process. Van Merriënboer's concept of "second-order scaffolding" — shared control over task selection (§0.21.1)● — provides the design mechanism: teacher and learner jointly plan the learning trajectory rather than the teacher dictating it. This is the autonomy-structure integration at its most natural — structure decreases as competence increases, and autonomy grows in proportion.

At **Stage 3**, productive failure temporarily increases autonomy — students generate and explore without teacher intervention — followed by structured instruction in the Assembly phase. The PS-I design resolves the tension within a single lesson: the exploration phase gives autonomy (the student chooses which approaches to try), the instruction phase gives structure (the teacher assembles canonical knowledge). Lemov's Culture of Error — "mistakes are a first, positive, and often critical step toward getting it right" (2021, §9.10.4)● — provides the relational condition that makes this autonomy productive rather than anxiety- provoking.

At **Stage 4**, the student directs learning with coaching. The power relationship shifts fundamentally — the learner has enough expertise to make meaningful choices, and the teacher's role is to respond to the learner's direction rather than set it. Schön's "willing suspension of disbelief" — the student temporarily accepts the coach's authority as a condition for learning, trusting that autonomy will be restored — describes a trust-based authority transfer that is qualitatively different from Stage 1 compliance (1987, §0.20.45)●. The student chooses to defer, and the choice is revocable.

At **Stage 5**, autonomy dominates. Structure is self-imposed through metacognitive self-regulation. The learner identifies their own weaknesses, designs their own practice, and seeks feedback selectively. This is consistent with W2-008's finding that self-regulation develops through warm, predictable environments — by Stage 5, the environmental conditions have done their work and self-regulation has become internalized.

9.2 THE COMPLIANCE CASCADE AS SYSTEMIC FAILURE

Kohn's critique deserves engagement beyond the philosophical level. The compliance cascade is the systemic failure mode: when institutional pressure for "results" propagates downward, teachers become controlling, trapping students at Stage 1 indefinitely — "dull and repetitive skills instruction" that "concentrates its curriculum on 'basic' skills" for disadvantaged students (Kohn, 1999, §6.4.3)●. Christodoulou (2014) makes the complementary argument from the other direction: progressive pedagogy that substitutes projects and discovery for knowledge also harms disadvantaged students by denying them "the vital knowledge they need to make sense of the world" (Verified direct). Both critiques are correct. The staged model resolves the apparent contradiction: explicit instruction provides the knowledge base (Stage 1–2), productive failure and guided inquiry develop transfer and judgment (Stages 3–4), and the relational-environmental conditions that W2-008 identifies must be maintained throughout. The failure mode is not any single instructional approach but institutional systems that freeze learners at one stage.

9.3 THE RELATIONAL-ENVIRONMENTAL PRECONDITION

W2-008's most consequential finding for instructional design: self-regulation develops primarily through warm, predictable environments rather than direct cognitive training. The Watts et al. (2018) reanalysis of the marshmallow test found that two-thirds of the predictive effect disappeared with controls for family environment — the child who delays gratification does so partly because they live in a trustworthy, predictable world (W2-008 §3.3). Sweller et al. (2019) note from the CLT side that "stress, emotions and uncertainty may restrict the capacity of working memory" (p. 285)●, and Blair and Raver (2014) showed that chronic unpredictable stress directly degrades self-regulatory systems (as reported in W2-008 §3.3).

The implication for the staged model: the relational- environmental conditions are not a separate design problem — they are a *precondition* for the cognitive mechanisms to operate. Productive failure requires a Culture of Error (Lemov, 2021, §9.10.4–5)●. No Opt Out requires "a foundation of trust" (Lemov, 2021, §10.6.8)●. Brousseau's didactical contract — "these (specific) habits of the teacher are expected by the student and the behaviour of the student is expected by the teacher" (1997, §5.2.3)● — is the relational context within which all instruction occurs. Without it, none of the approaches work.

9.4 WHAT REMAINS UNRESOLVED

The honest assessment: the autonomy-structure tension is *resolvable in principle* — the distinction between the content of instruction (constrained by CLT) and the context of instruction (where SDT operates) is conceptually clean. Whether teachers can reliably implement autonomy-supportive explicit instruction at scale — combining CLT-optimal content with SDT-optimal delivery — is an open empirical question. The practitioner community converges on something the research has not tested: Lemov's "warm-strict" integration treats warmth and structure as a single, trainable

skill (Practitioner Gap P1). Whether this combination can be trained at scale through professional development is the central unanswered question.

10 AI-MEDIATED INSTRUCTIONAL DESIGN

10.1 THE METACOGNITIVE LAZINESS PROBLEM

Fan et al. (2024) found that university students using ChatGPT improved task performance on writing but showed no gains in knowledge or transfer compared to controls (Abstract-verified). The mechanism — what the authors term "metacognitive laziness" — is precisely what the staged-instruction model predicts: when the AI provides the cognitive work, the learner's working memory is not engaged in schema construction. The AI is an infinitely patient Topaze machine — scaffolding that eliminates the learning target.

This finding connects to Brousseau's fundamental paradox: "everything that [the teacher] undertakes in order to make the student produce the behaviours that she expects tends to deprive this student of the necessary conditions for [...] learning" (1997, §1.2.2)•. The AI tutor that answers questions, generates explanations, and completes tasks is the ultimate instantiation of this paradox — maximally helpful, maximally learning-defeating.

10.2 WHAT THE STAGED MODEL IMPLIES FOR AI DESIGN

The staged-instruction model suggests specific design principles for AI-mediated learning:

At **Stage 1**, AI can effectively deliver worked examples, provide immediate feedback on practice problems, and adapt pacing to individual learners — the traditional intelligent tutoring system role. CLT's guidance-fading effect applies: the AI should reduce support as the learner demonstrates competence.

At **Stage 2**, the AI should shift from providing solutions to scaffolding process — asking "what would you try next?" rather than showing the next step. The completion-problem design translates naturally to AI interaction.

At **Stage 3**, the critical design decision: the AI must *withhold* help during the productive-failure phase. This is counter to the default design of current AI assistants, which are optimized for helpfulness. An AI tutor designed on PF principles would present a challenging problem, let the learner struggle and generate multiple approaches, and only then provide structured instruction connecting the learner's attempts to canonical knowledge. The AI's advantage — it can track exactly what the learner generated and tailor the instruction phase accordingly — could make it a better Assembly tool than a human teacher, if designed correctly.

At **Stages 4–5**, the AI's role shifts to what Schön describes as the "virtual world" — a practice environment where the learner can take consequential actions without irreversible consequences. AI-generated scenarios, simulated clients, virtual design studios could provide the varied experience with valid feedback that W2-009 identifies as necessary for judgment development.

10.3 THE AUTONOMY PROBLEM IN AI TUTORING

AI tutoring has a specific and underexplored relationship to the autonomy-structure tension. AI can provide structure continuously and adaptively — no human teacher can match an AI's ability to calibrate difficulty, provide immediate feedback, and track individual learner states

across thousands of practice attempts. But AI cannot (yet) provide the relational warmth that W2-008 identifies as a precondition for self-regulation development. The risk is a high-structure, low-warmth environment — precisely the compliance-cascade configuration that Kohn warns against.

The Jerrim et al. (2019) longitudinal PISA analysis adds a cautionary note from a different angle: students in high-inquiry environments scored lower on science tests in later years (Abstract-verified). This finding is often cited against inquiry, but its relevance to AI-mediated instruction is worth considering. The "inquiry" measured by PISA is largely unscaffolded — students reporting that they "are allowed to design their own experiments" and "spend time in the lab." If unscaffolded inquiry produces worse outcomes than structured instruction, unscaffolded AI interaction (where the learner directs the conversation) may produce similarly poor results. The staged model's implication: AI-mediated instruction needs the same expertise-adaptive calibration as human instruction, not a one-size-fits-all chatbot interaction.

Kasneci et al. (2023) surveyed the emerging ChatGPT-in-education literature and identified both opportunities and risks (Abstract-verified). The opportunities they highlight — personalized feedback, adaptive pacing, availability outside school hours — correspond to what the staged model prescribes at Stages 1–2: structured, feedback-rich, individually calibrated instruction. The risks — academic dishonesty, over- reliance, shallow engagement — correspond to the metacognitive laziness problem and the Topaze effect.

The staged model suggests a specific design architecture for AI-mediated instruction: AI is well-suited to Stages 1–2 (where the cognitive demands are well-specified and feedback can be algorithmically generated) and to specific support functions at later stages (scenario generation at Stage 4, feedback provision at Stage 5). The risks concentrate at Stage 3 — productive failure requires that the AI *resist the temptation to help*, which contradicts the default optimization of every current AI assistant — and at the relational level, where the warm, predictable environment that W2-008 identifies as a precondition for self-regulation development cannot be algorithmically generated. The design challenge is not whether AI can teach — it clearly can, within well-specified domains — but whether AI can create the conditions under which learners develop the self-regulation, judgment, and metacognitive awareness that constitute Layers 3–5 of the competence stack.

11

PRACTICAL IMPLICATIONS FOR CURRICULUM DESIGN

11.1 WHAT A CURRICULUM DESIGNER CAN CONFIDENTLY DO

Based on the evidence reviewed, the following prescriptions carry high confidence:

1. Design instruction as an expertise-adaptive sequence, not a fixed approach. The strongest finding in this review: the optimal instructional approach depends on the learner's current expertise in the domain. Novices need explicit instruction; developing learners benefit from productive failure and scaffolded inquiry; advanced learners benefit from independent practice and coaching. This is not "it depends" — it is a specific, evidence-grounded trajectory.

2. For novices in well-structured domains, use explicit instruction with worked examples. This is one of the most robust findings in education research. The worked-example effect, the guidance-fading effect, Rosenshine's principles, and Engelmann's examples/non-examples all converge.

3. For novices in ill-structured domains, use exemplar-based instruction with guided analysis. Direct instruction of rules does not work (Hillocks ES = .02). Exemplar analysis — studying high-quality models with structured questions — serves the role worked examples serve in well-structured domains.

4. Design productive-failure tasks at the transition from foundational knowledge to conceptual understanding. When learners have enough prior knowledge to generate relevant (if incorrect) approaches, the PS-I sequence produces up to 3x the effect of traditional I-PS.

5. Use the environmental mode for ill-structured skills. Structured problem-solving activities with clear objectives, specific materials, and peer interaction — neither direct instruction nor free exploration — are the most effective approach to teaching writing and likely other ill-structured skills.

6. Embed assessment in instruction. Retrieval practice at the start of each session (spacing and interleaving effects)○. Formative assessment through checking for understanding — Sherrington identifies this as "the single biggest common area for improvement" in teaching practice (2019, Strand 2)●. Cold Call rather than volunteer sampling for more accurate data on student understanding (though, as noted in Practitioner Gap P2, this has no controlled evidence). Low-stakes practice with immediate feedback. These are not add-ons but integral components — assessment IS instruction when designed correctly (supported by W2-003's findings on assessment and feedback)○.

7. Maintain the relational-environmental foundation throughout. This is the W2-008 integration point. Warmth, predictability, trust — these are the environmental conditions under which self-regulation develops (Watts et al., 2018, as reported in W2-008 §3.3). Lemov's Culture of Error provides the instructional operationalization: "mistakes are a first, positive, and often critical step toward getting it right" (2021, §9.10.4)●. No Opt Out works on "a foundation of trust" rather than compliance (§10.6.8)●. Brousseau's didactical contract — the set of mutual expectations between teacher and student — is the relational infrastructure within which all instruction operates. Without it, the cognitive mechanisms do not function: chronic unpredictable stress degrades the very working-memory capacity that CLT's prescriptions aim to optimize (Blair & Raver, 2014, as reported in W2-008 §3.3). The relational-environmental foundation is not a separate concern from instructional design — it is a precondition for instructional design to work.

11.2 WHAT A CURRICULUM DESIGNER SHOULD BE CAUTIOUS ABOUT

1. Generalizing from well-structured to ill-structured domains. The evidence base for most instructional-design prescriptions is from mathematics, science, and reading — well-structured domains with clear right answers and decomposable procedures. Extending to writing, design, ethics, and artistic practice requires the adapted approaches described in Section 8 — not direct transfer. The assumption that "what works in math will work in writing" is contradicted by Hillocks' evidence: direct instruction of rules (ES = .02 in writing) is one of the most effective approaches in mathematics. The staged model's ill-structured variant (Section 7) addresses this, but the caution remains: always check whether the evidence base for a recommendation includes the domain in question.

2. Applying PF without the design conditions. Productive failure requires carefully designed tasks (accessible, multiple solutions, layperson's language), a Culture of Error where mistakes are treated as learning opportunities, and explicit Assembly in the instruction phase where the teacher connects canonical knowledge to students' generated approaches. Poorly designed PF — assigning a hard problem and then lecturing without connecting to students' work — is just failure. The design conditions are load-bearing, and Kapur is explicit about this: the 3x effect size applies when the conditions are met, not when the label is applied.

3. Assuming 4C/ID can be implemented without design teams. The model's sophistication — task analysis, classification of recurrent and non-recurrent skills, design of task classes at multiple complexity levels, coordination of supportive and procedural information — requires instructional-design expertise that individual teachers typically lack. Van Merriënboer (2018) is explicit: "teacher design teams are strongly preferred." Simplified implementations are needed for K-12 contexts, and none currently exist with empirical validation.

4. Treating the transition signals as precise diagnostics. The transition from one stage to the next is a matter of teacher judgment informed by rough heuristics, not a precise measurement. The 80% threshold is one study's finding, not a universal law. The deep-structure recognition signal (Stage 3→4) is a theoretical description of what expert cognition looks like, not a classroom-usable test. A curriculum designer should specify the stages and the approximate transition criteria, but should also acknowledge that the actual transition depends on the teacher's professional judgment — and should invest in developing that judgment rather than attempting to eliminate it.

11.3 WHAT THE EVIDENCE DOES NOT YET TELL US

1. How to teach effectively in ill-structured domains beyond writing. The environmental mode has strong evidence for writing instruction. Whether it transfers to design, ethics, clinical reasoning, or artistic practice is unknown. The convergence of independent traditions (Section 8) is suggestive but not meta-analytically confirmed.

2. Whether instructional approach affects far transfer. Near transfer is achievable with well-designed instruction. Far transfer remains elusive regardless of approach.

3. Long-term motivational effects. Whether a curriculum built on the staged model sustains motivation better than alternatives is untested over semesters or years.

4. The interaction between AI tutoring and the staged model. Whether AI can effectively implement the withholding required at Stage 3 and the relational warmth required throughout is an open design question.

11.4 DOMAIN-SPECIFIC PRESCRIPTIONS

Well-structured domains (mathematics, science procedures, grammar, programming syntax): These are the domains where CLT's evidence base is strongest. Stage 1 uses worked examples, Rosenshine's principles, and Engelmann's examples/non-examples. Stage 3 uses productive failure in the PS-I format — Kapur's paradigm case (statistics) is the proof of concept. Stage 4 uses guided inquiry with open problems having multiple solution paths. The key design principle is the guidance-fading effect: reduce support systematically as schemas develop. The transition signals are relatively clear because performance can be objectively assessed.

Ill-structured domains (writing, design, ethical reasoning, artistic practice): These are the domains where the review's contribution is most significant, because they are the domains CLT's prescriptions do not address. Stage 1 is modified — direct instruction of rules does not work (Hillocks ES = .02). Instead, use exemplar analysis ("Follow me!" demonstration, model texts, think-aloud of expert reasoning). Stage 3 is the dominant mode, not a transitional one — the environmental mode (structured tasks + peer interaction + clear objectives) is where most instructional time should be spent. Stage 4 uses the design studio model (Schön's joint experimentation, reflective conversation). The key design principle is devolution, not fading: transfer responsibility for the problem itself, not just withdraw scaffolding from a fixed problem.

Mixed domains (clinical reasoning, engineering, teaching practice): These domains combine recurrent sub-skills (reading lab results, using measurement instruments, managing classroom logistics) with non-recurrent judgment tasks (diagnosing a patient, designing a structure, responding to student misconceptions). The 4C/ID model is most directly applicable: use task classes with whole tasks at increasing complexity. Case-based instruction with progressive complexity serves Stage 3. Clinical supervision and apprenticeship serve Stages 4–5. The key design principle is the recurrent/non-recurrent distinction: procedural information (just-in-time, then faded) for recurrent skills; supportive information (cognitive strategies, domain models, not faded but elaborated) for non-recurrent skills.

12

CLOSING ASSESSMENT: WHAT WE KNOW, WHAT WE DON'T, AND WHAT CHANGED

12.1 CONFIDENCE LEVELS

High confidence: - The expertise reversal effect is real and consequential (Tetzlaff et al. 2025 meta-analysis)○. - Unguided discovery is ineffective for novices (KSC 2006, multiple meta-analyses). - Guided inquiry is effective when properly scaffolded (Lazonder & Harmsen 2016, Hmelo-Silver et al. 2007). - Productive failure reliably enhances conceptual understanding and transfer (Kapur 2024 meta-analysis)●. - The outcome measure determines which approach "wins" (de Jong et al. 2023 convergence)○. - Direct instruction of rules fails in writing instruction (Hillocks 1986)●. - The DI-vs-inquiry debate has converged on an integrative position (de Jong 2023 + Sweller 2023 exchange)○.

Medium confidence: - The five-stage model captures the right structure, but transition signals are rough heuristics, not precise diagnostics. - The environmental mode (structured activity + coaching) is the right approach for ill-structured domains, but the evidence is strongest for writing and needs extension. - The autonomy-structure tension is resolvable in principle through autonomy-supportive delivery of structured content, but implementation at scale is untested. - 4C/ID is the most promising framework for complex learning but requires adaptation for K-12 contexts. - Productive failure works beyond mathematics, but the evidence base outside math is narrower.

Low confidence: - Whether the staged model's later stages (4–5) can be operationalized for classroom use. - Whether instructional approach affects far transfer meaningfully. - Long-term motivational effects of any instructional approach. - Cultural variation in optimal instructional approaches (almost all evidence is from Western contexts). - Whether AI tutoring designed on PF principles can avoid the metacognitive laziness problem.

12.2 WHAT V2 RESOLVED THAT V1 COULD NOT

1. **Primary-text engagement transformed the analysis.** v1 engaged Kirschner, Sweller, van Merriënboer, Kapur, Rosenshine, Engelmann, and Chi mostly from secondary sources and training knowledge. v2 reads the primary texts — fourteen of them — and discovers things the secondary accounts systematically miss. Engelmann's theoretical framework is logical, not behavioral — "faultless communication" is a theory of instructional sequence design, not a theory of reinforcement. Rosenshine's evidence base is 1970s–90s elementary math and reading, not the universal prescriptions the practitioner movement implies. Brousseau's fundamental paradox — that every act of teaching risks eliminating the learning it aims to produce — is a structural insight absent from the entire Anglo-American tradition. Hillocks' devastating .02 effect size for direct instruction of writing rules transforms a vague claim about ill-structured domains into a precise empirical finding.

2. **The staged-instruction model replaces the binary.** v1 described the trajectory abstractly — "direct instruction for novices, inquiry for experts." v2 specifies five stages with transition signals, worked examples for both well-structured (statistics/standard deviation) and ill-structured

(persuasive essay writing) domains, a competence-stack mapping, and explicit documentation of where the evidence is strong and where it runs thin. The model is concrete enough to guide instructional planning while honest enough to flag its own limitations.

3. **The ill-structured domain gap (v1 Gap 1) is substantially narrowed.** This was v1's most consequential gap and the most important for the lab's curriculum mission. v2 addresses it through direct engagement with five domain-specific traditions: writing pedagogy (Hillocks), design education (Schön), French *didactique* (Brousseau, Chevallard), clinical reasoning (Schmidt, Lubarsky), and the non-English traditions (Galperin, Klafki, lesson study). The convergence finding — three independent traditions, developed in different countries and different decades, independently discovering that structured activity with coaching works and direct instruction of rules fails in ill-structured domains — is the review's most important empirical contribution. Devolution as the ill-structured equivalent of scaffolding-fading is the review's most important conceptual contribution.

4. **The three-traditions map clarifies the field.** v1 presented instructional design as a two-sided debate. v2 identifies three traditions that rarely communicate — cognitive psychology, learning sciences, and instructional practice — and shows that the binary debate is partly an artifact of this non-communication. This reframing changes how we approach the literature: instead of asking "which side is right?" we ask "what does each tradition contribute that the others miss?"

5. **Non-English sources expand the theoretical base.** v1 cited exclusively English-language sources. v2 engages Brousseau and Chevallard (French *didactique*), Galperin (Russian activity theory, via Engeness 2020), Klafki (German *Didaktik*, via Sjöström & Eilks 2020), and Japanese lesson study (Takahashi & McDougal 2016). The French *didactique* engagement is particularly consequential — Brousseau's concepts (milieu, devolution, the Topaze/Jourdain effects, epistemological obstacles) provide diagnostic and design tools that the Anglo-American tradition lacks entirely.

6. **Practitioner gap detection identifies research opportunities.** Seven practitioner communities systematically assessed, yielding seven gaps the research literature has not investigated — including the warm-strict integration as a trainable skill (P1), Cold Call vs. volunteer sampling for formative assessment accuracy (P2), and DI program effects in ill-structured domains (P6). These are not speculative research questions — they are phenomena that practitioners have identified and operationalized that the academic literature has not tested.

12.3 WHAT REMAINS GENUINELY UNKNOWN

The honest accounting: several of v1's gaps remain open because the evidence does not exist to close them.

Expertise assessment for adaptive instruction (v1 Gap 2). How to diagnose where a learner sits on the expertise continuum for real-time instructional calibration remains the central practical problem. The staged model describes what instruction should look like at each stage, but it does not solve the prior problem: how does a teacher (or an AI tutor) determine which stage a given learner occupies for a given topic on a given day? Kalyuga and Sweller (2004) proposed rapid diagnostic tests, but these have seen limited classroom implementation (Training-derived). The 80% success-rate heuristic is a rough signal. Sherrington's honest answer — "a subtle skill, a central element of teacher expertise" — is the practitioner community's acknowledgment that this problem is unsolved. Technology-enhanced approaches (adaptive testing, learning analytics) may offer a path forward, but the integration of real-time expertise diagnosis with the staged model's prescriptions remains an engineering challenge as much as a scientific one.

Long-term motivational effects (v1 Gap 3). Whether the staged model sustains motivation better than fixed approaches over semesters and years — and whether the compliance cascade materializes in DI-heavy curricula — is untested. The current evidence base operates on a timescale of days to weeks. Curriculum operates on a timescale of years. This mismatch means that our most confident prescriptions (explicit instruction for novices, productive failure for conceptual understanding) are validated for short-term outcomes that may not predict long-term ones. The Stockard affective-outcomes exception is the strongest indirect evidence that short-term achievement and long-term motivation may diverge, but it is suggestive, not dispositive.

Transfer (v1 Gap 7). Whether instructional approach reliably affects far transfer remains genuinely uncertain. Productive failure shows transfer advantages within domains, and the "preparation for future learning" paradigm (Schwartz & Bransford, 1998) provides a theoretical framework. But cross-domain transfer — applying principles from one field to genuinely different problems in another — remains elusive regardless of approach. The implication for curriculum design is sobering: "general thinking skills" may be unreachable as a direct educational goal, and competencies may need to be developed within each domain rather than taught once and expected to transfer.

Cultural variation (v1 Gap 5). Almost all evidence comes from Western, educated, industrialized contexts. The non-English tradition engagement in this review (Brousseau, Galperin, Klafki, lesson study) shows that effective instruction can be theorized differently in different cultures, but the experimental validation still comes primarily from Western samples. Whether the expertise reversal effect, the worked-example effect, or the productive-failure benefit hold across educational cultures is an empirical question that the field has not systematically addressed.

12.4 METHODOLOGICAL HONESTY

Several claims in this review rest on thin evidence, and intellectual honesty requires flagging them:

- The transition signals for Stages 3→4 and 4→5 are theoretically grounded (Chi et al. 1981; Schön 1983) but not operationalized as classroom diagnostics. They are descriptions of what happens, not tools for detecting when it happens. - The ill-structured domain convergence rests on the writing evidence (Hillocks 1986, one meta-analysis from 1986, not replicated at that scope) and the design evidence (Schön, qualitative case studies, no controlled experiments). The convergence is genuine but the evidence base for each tradition is thinner than for well-structured instruction. - The AI-mediated instructional design section (Section 10) rests on a single study (Fan et al. 2024) and emerging literature. The connection to Brousseau's Topaze effect is the review's theoretical contribution, not an empirical finding. - The practitioner-identified gaps (Section in gaps-v2.md) are hypotheses derived from practitioner claims, not established research findings. Their value is as research questions, not as evidence. - Galperin's stepwise formation model is read through a single secondary source (Engeness, 2020); the primary Russian-language literature was not directly accessed. - The argument that the staged model integrates all three traditions is this review's synthetic claim, not a position endorsed by any of the cited researchers. Sweller, Kapur, and Engelmann would likely object to aspects of the integration.

This review represents the best synthesis the current evidence supports. Where the evidence is strong (worked examples for novices, productive failure for conceptual understanding, the failure of direct instruction in writing), the prescriptions are correspondingly strong. Where the evidence is thin (later- stage transition signals, ill-structured domain transfer, AI design), the prescriptions are flagged as provisional. The gap between what we know and what the curriculum needs is genuine, and closing it honestly — rather than papering over it with confident-sounding recommendations — is the most useful thing this review can do for the PI.

Review complete.

REFERENCES

Ahmadi, A., et al. (2023). A classification system for teachers' motivational behaviors recommended in self-determination theory interventions. *Journal of Educational Psychology*, 115(9), 1158–1176.

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2010). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, 103(1), 1–18.

Blair, C., & Raver, C. C. (2014). Closing the achievement gap through modification of neurocognitive and neuroendocrine function. *PLoS ONE*, 9(11), e112393.

Brousseau, G. (1997). *Theory of Didactical Situations in Mathematics*. Kluwer. [French, English translation by Balacheff et al.]

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4), 219–243.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5(2), 121–152.

Christodoulou, D. (2014). *Seven Myths About Education*. Routledge.

Clark, R. E., Kirschner, P. A., & Sweller, J. (2012). Putting students on the path to learning: The case for fully guided instruction. *American Educator*, 36(1), 6–11.

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

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.

Dochy, F., et al. (2003). Effects of problem-based learning: A meta-analysis. *Learning and Instruction*, 13(5), 533–568.

Engelmann, S., & Carnine, D. (1991). *Theory of Instruction: Principles and Applications* (rev. ed.). ADI Press.

Engeness, I. (2020). P. Y. Galperin's development of human mental activity: Lectures in educational psychology. *Cultural-Historical Psychology*, 16(4). [Russian tradition, via English]

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

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(4), 1340–1362.

Fiorella, L. (2023). Making sense of generative learning. *Educational Psychology Review*, 35, 50.

Freeman, S., et al. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410–8415.

Frèrejean, J., et al. (2019). [4C/ID implementation in higher education]. *Educational Psychology Review*.

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, 99(3), 445–476.

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record*, 111(9), 2055–2100.

Hillocks, G., Jr. (1986). *Research on Written Composition: New Directions for Teaching*. ERIC Clearinghouse on Reading and Communication Skills / NCTE.

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist*, 42(2), 99–107.

Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review*, 16(3), 235–266.

Jerrim, J., Oliver, M., & Sims, S. (2019). The relationship between inquiry-based teaching and students' achievement. New evidence from a longitudinal PISA study. *Learning and Instruction*, 61, 35–44.

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

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, 38(1), 23–31.

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

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

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, 41(2), 75–86.

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

Kohn, A. (1999). *The Schools Our Children Deserve: Moving Beyond Traditional Classrooms and "Tougher Standards."* Houghton Mifflin.

Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. *Review of Educational Research*, 86(3), 681–718.

Lemov, D. (2021). *Teach Like a Champion 3.0: 63 Techniques That Put Students on the Path to College*. Jossey-Bass/Wiley.

Lubarsky, S., Dory, V., Audétat, M.-C., Custers, E., & Charlin, B. (2015). Using script theory to cultivate illness script formation and clinical reasoning in health professions education. *Canadian Medical Education Journal*, 6(2), e61–e70.

van Merriënboer, J. J. G., & Kirschner, P. A. (2018). *Ten Steps to Complex Learning* (3rd ed.). Routledge.

Merrill, M. D. (2002). First principles of instruction. *Educational Technology Research and Development*, 50(3), 43–59.

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

Reeve, J., et al. (2004). [Pressure on teachers → controlling behavior]. Various publications.

Renkl, A., & Atkinson, R. K. (2003). Structuring the transition from example study to problem solving in cognitive skill acquisition: A cognitive load perspective. *Educational Psychologist*, 38(1), 15–22.

Rosenshine, B. (2012). Principles of instruction: Research-based strategies that all teachers should know. *American Educator*, 36(1), 12–39.

Sailer, M., et al. (2024). [ICAP boundary conditions]. *Educational Psychology Review*.

Schmidt, H. G., & Boshuizen, H. P. A. (1993). On the origin of intermediate effects in clinical case recall. *Memory & Cognition*, 21(3), 338–351.

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

Schön, D. A. (1987). *Educating the Reflective Practitioner: Toward a New Design for Teaching and Learning in the Professions*. Jossey-Bass.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction*, 16(4), 475–522.

Sherrington, T. (2019). *Rosenshine's Principles in Action*. John Catt Educational.

Sjöström, J., & Eilks, I. (2020). The Bildung theory — From von Humboldt to Klafki and beyond. In B. Akpan & T. J. Kennedy (Eds.), *Science Education in Theory and Practice*. Springer. [German tradition, via English]

Stockard, J., Wood, T. W., Coughlin, C., & Rasplica Khoury, C. (2018). The effectiveness of direct instruction curricula: A meta-analysis of a half century of research. *Review of Educational Research*, 88(4), 479–507.

Strobel, J., & van Barneveld, A. (2009). When is PBL more effective? *Interdisciplinary Journal of Problem-Based Learning*, 3(1).

Sweller, J. (2023). [Response to de Jong et al.]. *Educational Research Review*.

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31, 261–292.

Takahashi, A., & McDougal, T. (2016). Collaborative lesson research: Maximizing the impact of lesson study. *ZDM Mathematics Education*, 48, 513–526.

Tetzlaff, L., et al. (2025). [Expertise reversal effect meta-analysis]. *Educational Psychology Review*.

Vansteenkiste, M., et al. (2020). Basic psychological need theory: Advancements, critical themes, and future directions. *Motivation and Emotion*, 44, 1–31.

Walker, A., & Leary, H. (2009). A problem based learning meta analysis. *Interdisciplinary Journal of Problem-Based Learning*, 3(1).

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.