HOW SHOULD INSTRUCTION BE DESIGNED?

The Evidence Beyond the Binary

Applied Pedagogy Research Lab

Guido Bartolucci, Principal Investigator

guido@appliedpedagogy.com

LAB.APPLIEDPEDAGOGY.COM

L1-004 · March 2026

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

CONTENTS

Part I

THE DEBATE

1

THE QUESTION THAT WON'T GO AWAY

How should a teacher teach? The question is as old as education itself — Socrates asked it before he drank the hemlock — but its modern form was crystallized in 2006, when Paul Kirschner, John Sweller, and Richard Clark published a paper with a title designed to provoke: "Why Minimal Guidance During Instruction Does Not Work" (Kirschner, Sweller & Clark, 2006). That paper, which has now accumulated over 6,500 citations and a field-weighted citation impact of 137.97, argued that constructivist, discovery-based, problem-based, experiential, and inquiry-based teaching all share a common flaw: they overload the limited capacity of working memory by asking novice learners to simultaneously discover solutions and learn from the discovery process. The cognitive architecture of human beings, the authors argued, makes this an impossible demand.

The paper ignited a firestorm. Within a year, three major rebuttals appeared in the same journal. Hmelo-Silver, Duncan, and Chinn (2007) argued that well-designed problem-based and inquiry learning are not "minimally guided" at all — they involve substantial scaffolding, structured problems, and expert facilitation. Schmidt, Loyens, van Gog, and Paas (2007) argued that problem-based learning is actually compatible with human cognitive architecture because the problem context activates prior knowledge and provides a framework for schema construction. Kuhn (2007) raised a different objection entirely: direct instruction is efficient for knowledge transmission, but education aims at more than knowledge transmission. If we want students to develop the capacity for inquiry — to generate questions, evaluate evidence, and construct arguments — then they must practice inquiry. You cannot learn to swim by watching demonstrations on land.

Nearly two decades later, the debate continues. As recently as 2022, Zhang, Kirschner, Cobern, and Sweller reasserted the superiority of direct instruction (Zhang et al., 2022), prompting a reply from thirteen leading researchers on the inquiry side — including Hmelo-Silver, de Jong, Koedinger, and Linn — who argued that "a more complete and correct interpretation of the literature demonstrates that inquiry-based instruction produces better overall results for acquiring conceptual knowledge than does direct instruction" and that "a combination of inquiry and direct instruction may often be the best approach to support student learning" (de Jong et al., 2023).

This investigation takes the position that the binary framing — direct instruction versus inquiry — is the problem, not the answer. The evidence, when read carefully, does not support the superiority of either approach in isolation. What it supports is something more interesting and more useful: an expertise-adaptive model of instruction in which the mode of teaching should change as learners develop. The optimal instructional approach for a given learner at a given moment depends on their prior knowledge, the nature of the learning goals, the structure of the domain, and the kind of understanding being sought. The job of this dissertation is to map exactly when and how.

This matters for Applied Pedagogy because instructional design is where learning science meets practice. The cognitive foundations surveyed in the Lo investigation — cognitive load theory, working memory limitations, schema construction, the expertise reversal effect — are not interesting in themselves. They are interesting because they constrain and enable the design of instruction. The motivation findings from L1-002 — that autonomy support is essential for sustained engagement but that self-regulation must be taught explicitly — create a tension that instructional design must resolve. The assessment findings from L1-003 — that retrieval practice is one of the most robust learning tools available and that feedback design matters enormously — must

be integrated into instructional design as core components, not afterthoughts. This investigation sits at the intersection of everything the lab has learned so far.

Five questions guide the investigation:

1. What does the meta-analytic evidence show for direct instruction versus inquiry-based approaches across different outcomes?
2. Under what conditions does problem-based learning outperform traditional instruction, and vice versa?
3. How should instruction transition from more explicit to more open-ended as learners develop expertise?
4. What does the 4C/ID model offer for complex learning task design?
5. What is the evidence for specific instructional design frameworks — Rosenshine's Principles, Merrill's First Principles, Engelmann's Direct Instruction, and Chi's ICAP framework?

The central argument that will emerge is this: for novices in well-structured domains, explicit instruction with worked examples is the most efficient path to initial competence. As expertise develops, instruction should progressively shift toward guided inquiry, productive failure, and independent problem-solving. The transition is not a binary switch but a gradual fading of scaffolding calibrated to the learner's developing schema. And the whole picture is complicated — importantly so — by the fact that the optimal approach depends not just on expertise level but on what kind of learning outcome is being pursued, what kind of domain is being taught, and whether we are optimizing for short-term performance or long-term understanding and transfer.

2 THE COGNITIVE SCIENCE FOUNDATION

Before the debate can be properly evaluated, we need to understand what cognitive science has established about how learning works at the architectural level. Three findings are foundational: the severe limitations of working memory, the central role of schema construction in expertise, and the conditions under which cognitive load helps or hinders learning.

2.1 WORKING MEMORY: THE BOTTLENECK

Human working memory can hold approximately four chunks of novel information simultaneously (Cowan, 2001). This is not a cultural artifact or a pedagogical inconvenience — it is a structural feature of human cognitive architecture. When learners encounter new material, they must process it through this bottleneck. If the material exceeds working memory capacity, learning fails. This is not a failure of effort or motivation; it is a failure of cognitive architecture to accommodate the demand.

Long-term memory, by contrast, has no known capacity limits. Expertise consists largely of well-organized schemas stored in long-term memory that allow the expert to treat complex patterns as single chunks. A chess master does not see 32 individual pieces on a board; she sees familiar configurations — patterns she has stored over years of practice — and can reason about them as units. A skilled reader does not process individual letters; she recognizes words and phrases as chunks, freeing working memory for comprehension. 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.

2.2 COGNITIVE LOAD THEORY: THE FRAMEWORK

Cognitive load theory (CLT), developed by John Sweller and colleagues over four decades, provides the most comprehensive framework for understanding how instruction interacts with cognitive architecture (Sweller, 1994; Sweller, van Merriënboer & Paas, 1998, 2019). CLT distinguishes two types of cognitive load that matter for instructional design:

Intrinsic load is determined by the inherent complexity of the material being learned — specifically, by the number of elements that must be processed simultaneously (element interactivity). Learning that the chemical symbol for hydrogen is H imposes low intrinsic load because there is a single association. Understanding how photosynthesis works imposes high intrinsic load because multiple interacting elements — light energy, chlorophyll, carbon dioxide, water, glucose, oxygen — must be understood in relation to each other.

Extraneous load is imposed by poor instructional design. When a textbook places a diagram on one page and its explanation on another, forcing the learner to mentally integrate spatially separated information, extraneous load increases without any benefit to learning. When a lecture includes irrelevant anecdotes, decorative images, or redundant text that duplicates spoken narration, extraneous load increases. The goal of instructional design, from a CLT perspective, is to minimize extraneous load and manage intrinsic load so that working memory capacity is devoted to productive schema construction.

The 2019 retrospective by Sweller, van Merriënboer, and Paas — which has accumulated 1,740 citations and an FWCI of 106.32 — added an evolutionary dimension to the theory, 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). This distinction, while debated at the boundaries, has an important implication: for biologically secondary knowledge, which is what education is primarily concerned with, explicit instruction is not just one option among many — it is the mechanism through which such knowledge is typically acquired. The question is not whether explicit instruction is needed, but how and when it should be deployed.

CLT has also generated a catalog of specific instructional effects, each grounded in experimental evidence. The split-attention effect shows that when learners must mentally integrate information from separate sources (a diagram on one page, its explanation on another), extraneous load increases and learning decreases; physical integration of text and graphics outperforms separated presentation (Chandler & Sweller, 1991). The redundancy effect shows that presenting identical information in multiple forms (spoken narration plus identical on-screen text) can actually harm learning, because processing the redundant source consumes working memory without adding new information. The modality effect shows that presenting information through both visual and auditory channels (diagrams with narration rather than diagrams with text) can reduce load by distributing processing across two working memory subsystems.

Mayer (2002) — with an extraordinary field-weighted citation impact of 543.71, the highest of any source found in the lab's research — systematized these and other findings into principles of multimedia learning: coherence (remove extraneous material), signaling (highlight essential information), spatial contiguity (place related text and graphics near each other), temporal contiguity (present corresponding narration and animation simultaneously), segmenting (break complex lessons into learner-paced segments), and pre-training (teach key concepts before the main lesson). These principles are directly applicable to any curriculum that uses visual or multimedia materials — which is to say, any modern curriculum. They represent the most practically actionable output of CLT for everyday instructional design.

De Jong (2009) offered a thoughtful critique of CLT that should be acknowledged. He questioned the precision of cognitive load measurement (how do we know whether load is intrinsic or extraneous?), the stability of the tripartite distinction (Sweller himself later collapsed germane load into intrinsic load), and the theory's applicability to complex, open-ended learning tasks where intrinsic load is inherently high and cannot be easily decomposed. These criticisms do not invalidate CLT's core insights — the working memory bottleneck is real, and extraneous load is real — but they do suggest that the theory is most powerful as a set of design heuristics for well-structured content and less powerful as a predictive theory for complex, authentic learning.

2.3 THE WORKED EXAMPLE EFFECT

One of CLT's most robust findings is the worked example effect: novice learners who study worked examples — step-by-step demonstrations of how to solve a problem — learn more efficiently than novices who attempt to solve equivalent problems on their own (Sweller & Cooper, 1985; Atkinson, Derry, Renkl & Wortham, 2000). The mechanism is straightforward. When novices attempt to solve problems without sufficient schemas, they resort to means-ends analysis — a general problem-solving strategy that involves working backwards from the goal, identifying the current state, and searching for operators that reduce the difference. Means-ends analysis imposes heavy cognitive load because the learner must simultaneously track the goal state, the current state,

the differences between them, and the available operators. All of this load goes to problem-solving, not to learning.

Worked examples eliminate this burden. By showing the learner the solution path, they free working memory to attend to the structure of the solution — the underlying principles, the relationships between steps, the reasons why particular operations are applied. This attention to structure is what builds schemas. The worked example effect has been replicated across dozens of studies in mathematics, physics, computer programming, and other well-structured domains. It is one of the most robust findings in instructional design research.

But the worked example effect has a crucial boundary condition, and understanding that boundary condition is the key to resolving the instruction-inquiry debate.

2.4 THE EXPERTISE REVERSAL EFFECT

Kalyuga, Ayres, Chandler, and Sweller (2003) demonstrated that instructional techniques optimal for novices can become ineffective or even harmful as learners gain expertise. This is the expertise reversal effect, and it has accumulated 1,808 citations. The mechanism is elegant: as learners develop schemas in a domain, the information that was once helpful (detailed worked examples, step-by-step guidance) becomes redundant with their existing knowledge. Processing this redundant information imposes extraneous load — the learner must attend to guidance they no longer need, and this attention is wasted or, worse, interferes with the schemas they have already constructed.

The practical implication is profound: instruction must change as the learner develops. What is optimal for a novice is suboptimal for an intermediate learner and potentially counterproductive for an advanced learner. The worked example effect flips. Where novices benefit from complete worked examples, more advanced learners benefit from completion problems (partially worked examples that the learner must finish), and still more advanced learners benefit from unguided problem-solving. The optimal instructional approach is not a fixed point but a moving target that must be calibrated to the learner's current expertise.

Kalyuga (2007) spelled out the practical implications: instructors need diagnostic tools to assess learner expertise in real time, and instruction should be designed as a fading sequence — from full guidance to partial guidance to no guidance — with the transition points determined by the learner's demonstrated competence, not by a predetermined schedule. This is easier said than done, and the practical challenges of implementing expertise-adaptive instruction at scale are considerable. But the principle is clear.

The expertise reversal effect, combined with the worked example effect, provides the cognitive science foundation for the resolution of the instruction-inquiry debate. Explicit instruction is not always better than inquiry, and inquiry is not always better than explicit instruction. Each is optimal at a different point in the learner's development, for different kinds of material, and for different learning goals. The debate is not about which is correct. It is about when each is appropriate.

3 THE DEBATE

3.1 THE ORIGINAL ARGUMENT

Kirschner, Sweller, and Clark (2006) is the most-cited paper in instructional design in the last two decades. Understanding what it actually argues — as opposed to what it is often caricatured as arguing — is essential.

The paper's core claim is that "minimally guided instruction" is less effective and less efficient than "instructional approaches that place a strong emphasis on guidance of the student learning process." The evidence cited comes from CLT, from the worked example research, and from studies comparing guided and unguided instruction across multiple domains. The argument rests on the working memory bottleneck: novice learners who must simultaneously search for solutions and learn from the search process face an impossible cognitive load.

Crucially, the paper defines "minimal guidance" as instruction in which "learners, rather than being presented with essential information, must discover or construct essential information for themselves." Under this definition, pure discovery learning — in which learners explore a domain with no structure, no feedback, and no direction — is the target. The paper lumps together constructivist teaching, discovery learning, problem-based learning, experiential learning, and inquiry-based teaching as all sharing this flaw.

This lumping is where the paper is most vulnerable to criticism, and where the subsequent debate has been most productive.

3.2 THE HMELO-SILVER REBUTTAL: PBL IS NOT MINIMALLY GUIDED

Hmelo-Silver, Duncan, and Chinn (2007) — a paper that has accumulated 2,446 citations — argued that Kirschner et al. fundamentally mischaracterized problem-based learning and inquiry learning. Well-designed PBL, they pointed out, is not "minimally guided." It involves:

- **Structured problems** that have been carefully designed to target specific learning objectives, not random exploration of a domain.
- **Scaffolding** — explicit support structures that guide learners through the problem-solving process, including prompts, hints, worked examples at critical junctures, and structured reflection.
- **Facilitation** — an expert teacher who actively guides student thinking, asks probing questions, redirects unproductive lines of inquiry, and provides direct instruction when needed.
- **Fading** — the gradual withdrawal of scaffolding and facilitation as students develop competence.

The key distinction is between unguided discovery (which the evidence does show is ineffective for novices) and scaffolded inquiry (which the evidence shows can be highly effective). Hmelo-Silver et al. argued that Kirschner et al. attacked a straw man: nobody in the PBL community advocates for unstructured, unsupported exploration as an instructional method. 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, not because inquiry as such is flawed. As Hmelo-Silver (2004) showed in her review of PBL's cognitive mechanisms, the approach works by activating prior knowledge, supporting knowledge construction through collaborative problem-solving, and developing self-directed learning skills — all of which are consistent with CLT's emphasis on schema construction.

3.3 THE SCHMIDT REBUTTAL: PBL AND COGNITIVE ARCHITECTURE

Schmidt, Loyens, van Gog, and Paas (2007) made a different argument: PBL is not just not-minimally-guided, it is actually compatible with human cognitive architecture. Their reasoning is that the problem context in PBL serves as an advance organizer — it activates relevant prior knowledge and provides a meaningful framework for organizing new information. Rather than increasing extraneous load, a well-designed problem can reduce it by providing context that makes new information meaningful. Learning that force equals mass times acceleration is an abstraction that imposes load. Learning it in the context of figuring out why a heavier car is harder to stop provides a schema-building context that supports the same learning.

This argument has merit but requires qualification. It works when learners have sufficient prior knowledge to make the problem context meaningful — that is, when the problem activates existing schemas rather than demanding the construction of entirely new ones. For genuine novices with no relevant prior knowledge, the problem context may add load rather than reducing it, because the learner must simultaneously understand the context and the content. This is precisely the expertise-dependence that the broader evidence supports.

3.4 THE KUHN OBJECTION: WHAT ARE WE TRYING TO TEACH?

Kuhn (2007) raised the most philosophically important objection. She argued that the debate about instructional effectiveness cannot be resolved without first asking: effective for what? Direct instruction may be more efficient for transmitting specific knowledge and procedures. But if the educational goal includes developing the capacity for inquiry — the ability to generate questions, design investigations, evaluate evidence, and construct arguments — then some form of inquiry-based instruction is necessary, because these are skills that must be practiced, not merely described.

This objection cuts deep because it reveals that the debate is partly about values, not just about evidence. If you define "effectiveness" as performance on a content knowledge test administered shortly after instruction, direct instruction usually wins. If you define it as the ability to solve novel problems, apply knowledge in new contexts, or sustain learning independently, the picture changes. And if you define it as the development of inquiry skills themselves — the capacity for scientific thinking, evidence evaluation, and argument construction — then inquiry-based instruction becomes not just a means to an end but a necessary component of the curriculum.

The Kuhn objection does not invalidate CLT or the worked example effect. It does, however, invalidate the claim that the debate can be settled by comparing test scores on knowledge assessments alone. The outcome measures matter, and they matter a lot.

4 THE META-ANALYTIC EVIDENCE

The debate has generated a substantial body of meta-analytic evidence. Reading it carefully — attending to what outcomes are measured, how "direct instruction" and "inquiry" are defined, and what moderators are examined — is essential to forming an honest assessment.

4.1 PBL META-ANALYSES

Dochy, Segers, Van den Bossche, and Gijbels (2003) conducted an early meta-analysis of PBL studies, primarily in medical and professional education. They found a robust positive effect of PBL on skills (applying knowledge, clinical reasoning) but a less consistent and sometimes negative effect on knowledge as measured by conventional tests. The FWCI of 125.54 reflects the paper's substantial influence.

Strobel and van Barneveld (2009) performed a meta-synthesis of existing PBL meta-analyses and arrived at a clear pattern: "PBL was superior when it comes to long-term retention, skill development and satisfaction of students and teachers, while traditional approaches were more effective for short-term retention as measured by standardized board exams." This finding — that the relative effectiveness of PBL depends on the outcome measure and the time horizon — has been the most consistent finding across the meta-analytic literature.

Walker and Leary (2009) conducted a more detailed meta-analysis that examined moderators. They found that the effectiveness of PBL varied by problem type, implementation type, discipline, and assessment level. PBL effects were larger for assessments that measured application and problem-solving than for assessments that measured factual recall. Effects were larger in medical education (where PBL has the longest implementation history and the most refined design) than in other disciplines.

4.2 INQUIRY LEARNING META-ANALYSIS

Lazonder and Harmsen (2016) conducted a comprehensive meta-analysis of inquiry-based learning — 72 studies — that explicitly examined the role of guidance. Their findings, published in the *Review of Educational Research* with a remarkable FWCI of 209.59, showed:

- A substantial overall effect of guidance on learning outcomes ($d = 0.50$).
- That guided inquiry was more effective than unguided inquiry across all outcome measures.
- That the type of guidance mattered for some outcomes but not others.

The key insight from this meta-analysis is that the question is not "inquiry versus direct instruction" but "inquiry with what kind of guidance?" When inquiry is properly scaffolded, it produces substantial learning gains. When it is unscaffolded, it does not. The Kirschner et al. critique of "minimal guidance" is valid — minimally guided inquiry is indeed ineffective. But this is an argument for better guidance, not for the abandonment of inquiry.

The pattern across PBL meta-analyses is striking in its consistency: 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 is not a neutral choice. It determines which approach "wins."

This pattern has a straightforward explanation grounded in CLT and schema theory. Direct instruction is optimized for efficient schema construction — it delivers information in a structured, organized form that minimizes extraneous load. This produces rapid knowledge acquisition, which is reflected in immediate post-tests. PBL, by contrast, requires learners to construct their own organizational framework for the knowledge they acquire. This is more effortful and less efficient in the short term, but the resulting schemas may be more deeply connected, more flexibly organized, and more connected to problem-solving contexts — qualities that support long-term retention and transfer. The inefficiency of PBL in the short term may be the desirable difficulty that produces superior long-term outcomes.

Hmelo-Silver (2004) reviewed the cognitive mechanisms through which PBL produces its effects. The problem-solving context activates prior knowledge (giving new information something to connect to), collaborative discussion elaborates understanding (requiring learners to articulate and defend their reasoning), self-directed learning develops metacognitive skills (learners must identify what they need to know), and the authentic problem context provides meaningful structure that supports schema organization. These mechanisms are all consistent with cognitive science — they are not mysterious. What is distinctive about PBL is that it deploys them simultaneously through a single instructional design, rather than treating them as separate interventions.

4.3 THE DE JONG ET AL. CONVERGENCE

In 2023, a landmark paper appeared that may represent the closest thing to a resolution the field has produced. Thirteen leading researchers from the inquiry tradition — including Ton de Jong, Ard Lazonder, Clark Chinn, Frank Fischer, Cindy Hmelo-Silver, Kenneth Koedinger, and Marcia Linn — jointly authored "Let's talk evidence: The case for combining inquiry-based and direct instruction" (de Jong et al., 2023). Their conclusion:

> Inquiry-based instruction produces better overall results for acquiring conceptual knowledge than does direct instruction… inquiry-based and direct instruction each have their specific virtues and disadvantages and… the effectiveness of each approach depends on moderating factors such as the learning goal, the domain involved, and students' prior knowledge and other student characteristics… a combination of inquiry and direct instruction may often be the best approach to support student learning.

This paper — with an FWCI of 54.65, already substantial for a 2023 publication — represents a significant shift from the polemics of the earlier debate toward evidence-based nuance. The key claims are:

1. Inquiry-based instruction produces better conceptual knowledge outcomes than direct instruction when guided adequately.
2. Direct instruction is more efficient for teaching procedures and factual knowledge.
3. The optimal approach combines both, with the balance depending on learning goals, domain, and learner characteristics — especially prior knowledge.
4. Guidance during inquiry can and often should include direct instruction at appropriate junctures.

This convergence does not mean the debate is over — the Sweller camp has not endorsed this position, and genuine disagreements remain about how to weight different outcome measures. But the center of gravity in the field has shifted decisively toward an integrative position.

4.4 WHAT THE META-ANALYSES CONVERGE ON

Across the meta-analytic literature, five findings emerge consistently:

1. **Unguided discovery is ineffective for novices.** Kirschner et al. were right about this. Purely unstructured exploration, without scaffolding or feedback, does not produce effective learning for learners who lack relevant prior knowledge. The cognitive load is too high.

2. **Guided inquiry is effective, often more effective than direct instruction for conceptual understanding.** When inquiry is properly scaffolded — with structured problems, expert facilitation, and calibrated support — it produces learning outcomes that equal or exceed direct instruction, particularly for conceptual understanding, application, and transfer.

3. **Direct instruction is more efficient for procedural knowledge and short-term recall.** When the goal is to transmit specific procedures or facts, direct instruction is faster and more reliable. There is no need to have students "discover" the quadratic formula.

4. **The outcome measure matters enormously.** Studies that measure factual recall tend to favor direct instruction. Studies that measure conceptual understanding, problem-solving, and transfer tend to favor guided inquiry. Studies that measure long-term retention tend to favor inquiry over direct instruction. The choice of outcome measure is not a neutral methodological decision — it embodies values about what counts as learning.

5. **Prior knowledge is the critical moderator.** The optimal instructional approach depends on what the learner already knows. Novices benefit from more guidance; advanced learners benefit from more independence. This is the expertise reversal effect operating at the level of instructional design.

Part II

BEYOND THE BINARY

5 PRODUCTIVE FAILURE

The most provocative challenge to the conventional sequence of instruction — teach first, then practice — comes from Manu Kapur's program of research on productive failure. Where CLT emphasizes the importance of minimizing unproductive cognitive load for novices, productive failure argues that a specific kind of struggle — attempting to solve problems before receiving instruction — can produce deeper learning than the conventional teach-then-practice sequence.

5.1 THE BASIC FINDING

Kapur's foundational studies (Kapur, 2008, 2009, 2010; Kapur & Bielaczyc, 2012) compared two instructional sequences:

- **Direct Instruction (DI):** Teacher explains the concept, demonstrates the procedure, students practice with feedback.
- **Productive Failure (PF):** Students attempt to solve a complex problem (which they cannot yet solve correctly), then receive instruction that builds on their attempts.

The finding, replicated across multiple studies in mathematics: both sequences produced equivalent procedural knowledge — students could execute the steps equally well. But productive failure produced significantly better conceptual understanding and transfer. Students who struggled first and received instruction second understood the concept more deeply and could apply it to novel problems more flexibly.

As Kapur reports in his 2024 book *Productive Failure*, a meta-analysis of over 50 studies with more than 160 comparisons confirms this pattern: productive failure reliably produces gains in conceptual understanding and transfer, sometimes as large as two academic years' worth of additional learning, while matching direct instruction on procedural outcomes (Kapur, 2024).

5.2 WHY DOES IT WORK? THE FOUR MECHANISMS

Kapur (2024) identifies four mechanisms — what he calls the "4A's" — that explain why productive failure works:

Activation. The initial problem-solving attempt activates learners' prior knowledge — even incomplete, incorrect prior knowledge. This activation creates cognitive hooks that subsequent instruction can attach to. Without this activation, instruction enters a cognitive vacuum — the learner has no existing schemas to connect the new information to, and the instruction, however clear, may fail to stick.

Awareness. Attempting and failing at a problem makes learners aware of what they do not know. This metacognitive awareness — the recognition of a knowledge gap — creates a readiness for instruction that passive listening does not. As VanLehn et al.'s research on tutoring dialogues showed, learning gains are concentrated at moments of "impasse" — when students realize they are stuck. Without the impasse, instruction has nothing to address.

Affect. The struggle creates an emotional investment in the problem. The Zeigarnik effect — the tendency to remember unfinished tasks better than completed ones — keeps the problem active in

the learner's mind. Curiosity generated by the unresolved problem creates a motivational drive toward the subsequent instruction. The learner is not passively receiving information; they are actively seeking resolution to a problem they have engaged with personally.

Assembly. The instruction phase does not merely deliver information — it "assembles" the learner's prior attempts into a coherent understanding. The teacher compares and contrasts the learner's solutions (including incorrect ones) with the canonical solution, showing which features of the learner's approach were on the right track and which were not. This comparison-and-contrast process is more effective than presenting the solution cold, because the learner has a rich set of personal experiences to compare against.

5.3 THE INTELLECTUAL LINEAGE: DESIRABLE DIFFICULTIES AND PREPARATION FOR FUTURE LEARNING

Productive failure did not emerge in a vacuum. It belongs to a broader family of findings in the learning sciences that challenge the intuition that making learning easier makes learning better.

The concept of "desirable difficulties" — coined by Robert and Elizabeth Bjork — captures the paradox: conditions that make initial learning harder (spacing, interleaving, retrieval practice, generation) often produce better long-term retention and transfer than conditions that make initial learning feel easy (massing, blocking, rereading, passive review). The testing effect, which L1-003 documented as one of the most robust findings in cognitive psychology, is a desirable difficulty — effortful retrieval during a test strengthens memory traces more than additional study. Spacing is a desirable difficulty — distributing practice over time feels less effective during learning but produces dramatically better long-term retention.

Productive failure is, in this framing, a desirable difficulty applied to conceptual learning. The struggle phase is harder and feels less productive than receiving a clear explanation. But the struggle creates conditions that make the subsequent instruction more effective — just as effortful retrieval strengthens memory traces that passive review does not.

A crucial predecessor is Schwartz and Bransford's (1998) work on "preparation for future learning." They showed that analyzing contrasting cases before a lecture produced better understanding than either the lecture alone or the contrasting cases alone. The mechanism was not that students learned from the contrasting cases per se, but that analyzing the cases prepared them to learn more deeply from the subsequent lecture. The lecture was the vehicle for learning; the contrasting cases were the preparation that made the vehicle effective. This is exactly the mechanism that Kapur later operationalized as productive failure, and it provides the clearest theoretical bridge between CLT (which emphasizes minimizing unproductive load) and productive failure (which deliberately introduces a specific kind of productive load).

The connection to the generation effect and the pretesting effect is also important. Attempting to generate an answer — even when the attempt fails — produces stronger subsequent learning than passively studying the answer. Pan et al. (2020) showed that pretesting before video lectures reduced mind wandering and increased attention; St. Hilaire et al. (2020) showed that pretesting improved learning only when students used the pretest questions to guide their attention during the subsequent lecture. These findings, drawn from the Kapur book's synthesis, suggest that the mechanisms of productive failure operate at multiple cognitive levels: activation of prior knowledge, awareness of knowledge gaps, focused attention during instruction, and deeper encoding through generation and contrast.

5.4 IS PRODUCTIVE FAILURE IN TENSION WITH CLT?

On the surface, productive failure appears to contradict CLT. If novices have limited working memory and worked examples are optimal for novices, how can asking novices to solve problems they cannot yet solve be beneficial? Is productive failure not precisely the kind of "minimally guided instruction" that Kirschner et al. warned against?

The resolution lies in recognizing that productive failure is not minimally guided instruction. It is a two-phase instructional design in which the struggle phase is followed by explicit instruction. The struggle phase is not the end of the pedagogical story — it is the beginning. The mechanism is not that novices learn through discovery; it is that the struggle phase prepares novices to learn more deeply from the subsequent explicit instruction. As Loibl, Roll, and Rummel (2017) argued in their theoretical synthesis, the benefit of problem-solving before instruction lies in its preparation effects — activating prior knowledge, generating awareness of knowledge gaps, and creating cognitive structures that make subsequent instruction more meaningful.

This distinction is subtle but crucial. Productive failure does not claim that struggle alone produces learning. It claims that struggle followed by instruction produces deeper learning than instruction alone. The instruction is still essential. The difference is in the sequencing: struggle → instruction rather than instruction → practice. And the mechanism is not discovery learning — the students typically do not discover the correct solution during the struggle phase. The mechanism is cognitive preparation: the struggle creates the conditions for the instruction to be maximally effective.

Kapur himself acknowledges this: "The key insight is that the productive failure approach is not about learning from failure per se. It is about what happens after the failure — the instruction that follows" (Kapur, 2024). The instructional phase in productive failure typically involves explicit, direct instruction — the teacher explains the canonical concept and procedure. What makes it more effective than instruction without prior struggle is that the learner is now prepared to receive it.

5.5 BOUNDARY CONDITIONS

The evidence for productive failure is substantial, but it comes with important boundary conditions:

Domain. Most productive failure research has been conducted in mathematics, where problems with clear structures lend themselves to the generate-and-contrast approach. Evidence in other domains — science, language learning, motor skills — is growing but less extensive.

Task design. Productive failure does not work with arbitrary tasks. The struggle phase must be carefully designed so that learners can engage meaningfully with the problem — generating multiple representations and solution attempts — without the problem being so far beyond their capacity that they produce nothing useful. Kapur's design principles (accessible language, contextualized problems, multiple solution paths, contrasting cases) are essential, not optional.

The instruction phase. The instruction that follows the struggle must explicitly build on the students' attempts — comparing, contrasting, and refining them. Simply lecturing after the struggle, without reference to what students produced, eliminates the advantage. The assembly mechanism requires that the teacher use the students' work as the raw material for instruction.

Assessment conditions. Productive failure's advantages appear on measures of conceptual understanding and transfer, not on measures of procedural skill. If the only assessment is whether students can execute a procedure, productive failure offers no advantage over direct instruction — both produce equivalent procedural competence.

Motivational context. The struggle phase must occur in a psychologically safe environment where failure is normalized and expected. If students experience the struggle as high-stakes evaluation, the motivational damage (test anxiety, helplessness) may overwhelm the cognitive benefits. This connects directly to L1-002's finding that autonomy support and separation of practice from evaluation are essential for sustained motivation.

6 THE EXPERTISE REVERSAL EFFECT

The expertise reversal effect and productive failure, taken together, suggest a developmental trajectory for instruction that is more nuanced than either "always use direct instruction" or "always use inquiry." The evidence points toward a sequence that evolves as the learner develops:

6.1 STAGE 1: NOVICE — FULL GUIDANCE

For genuine novices encountering a new domain, the evidence strongly favors explicit instruction with worked examples. The learner has no relevant schemas to activate, and working memory capacity is needed for schema construction. Unguided exploration at this stage is likely to produce frustration without learning.

However — and this is the productive failure insight — even novices may benefit from a brief period of structured struggle before instruction, if the struggle is carefully designed to activate whatever prior knowledge does exist and to create awareness of the knowledge gaps that instruction will address. The key is that this struggle phase is short, structured, and followed immediately by explicit instruction. It is not extended discovery learning; it is cognitive preparation for direct instruction.

6.2 STAGE 2: ADVANCED BEGINNER — FADED SCAFFOLDING

As learners develop initial schemas, the evidence supports a gradual fading of scaffolding. Complete worked examples give way to completion problems (partially worked examples that the learner must finish), which in turn give way to increasingly open problems. The Renkl, Atkinson, and colleagues' research on fading (Renkl, Atkinson, Maier & Staley, 2002) showed that this gradual transition produces better learning than an abrupt shift from worked examples to unguided problem-solving. The learner is progressively taking on more of the cognitive work as their schemas develop to support it.

6.3 STAGE 3: INTERMEDIATE — PRODUCTIVE FAILURE AND GUIDED INQUIRY

At the intermediate level — where learners have sufficient domain knowledge to engage meaningfully with problems but have not yet achieved expertise — the evidence supports more open instructional approaches. Productive failure becomes increasingly powerful because learners have enough prior knowledge to generate meaningful (if incorrect) solutions, and the compare-and-contrast process during the instruction phase can build on a richer base of existing schemas. Guided inquiry becomes appropriate because learners can manage the cognitive load of exploring a problem space while still benefiting from scaffolding and feedback.

6.4 STAGE 4: ADVANCED — INDEPENDENCE AND EXPERTISE BUILDING

For advanced learners approaching expertise, the evidence from the expertise reversal effect argues for removing scaffolding entirely. Worked examples become redundant; guided inquiry becomes

constraining. At this stage, learners benefit from independent problem-solving, deliberate practice targeting specific weaknesses, and the kind of extended investigation that characterizes genuine expertise development. The teacher's role shifts from instructor to coach — providing targeted feedback, designing appropriately challenging problems, and supporting metacognitive reflection.

6.5 THE PRACTICAL CHALLENGE

This developmental trajectory is easy to describe and extraordinarily difficult to implement in practice. The fundamental practical challenge is assessment: how does a teacher know when a learner has moved from Stage 1 to Stage 2, or from Stage 2 to Stage 3? The expertise reversal literature acknowledges this challenge but offers limited practical guidance. Kalyuga and Sweller (2004) proposed using rapid diagnostic tests to assess learner expertise in real time, but classroom implementation of such assessment remains rare.

In heterogeneous classrooms — which is to say, nearly all classrooms — the challenge is compounded by the fact that different learners may be at different stages simultaneously. The worked example that benefits one student may be redundant for another. The productive failure task that challenges one student may be inaccessible to another. Differentiating instruction along the expertise continuum at scale is the central practical problem of instructional design, and the field has not solved it.

This is where the instructional design frameworks examined in the next section become valuable — not because they solve the problem of expertise-adaptive instruction, but because they provide structured approaches for managing the complexity.

Part III

FRAMEWORKS

7 PRACTICAL FRAMEWORKS

Several frameworks attempt to translate the research base into actionable design principles. They vary in their scope, evidence base, theoretical foundations, and practical usability. Understanding what each offers — and what each lacks — is essential for curriculum design.

7.1 ROSENSHINE'S PRINCIPLES OF INSTRUCTION

Barak Rosenshine's (2012) "Principles of Instruction: Research-Based Strategies That All Teachers Should Know" is arguably the most influential practitioner-oriented synthesis of instructional research. Published in *American Educator*, it distilled decades of research into 10 principles:

1. Begin a lesson with a short review of previous learning.
2. Present new material in small steps with student practice after each step.
3. Ask a large number of questions and check the responses of all students.
4. Provide models.
5. Guide student practice.
6. Check for student understanding.
7. Obtain a high success rate.
8. Provide scaffolds for difficult tasks.
9. Require and monitor independent practice.
10. Engage students in weekly and monthly review.

These principles are solidly grounded in the process-product research tradition — studies that identified what effective teachers actually do in classrooms where students learn the most. The principles align well with CLT: small steps reduce working memory overload; models provide worked examples; guided practice builds schemas; review leverages the spacing and testing effects.

Strengths. Rosenshine's principles are concrete, actionable, and backed by substantial evidence from classroom research. They are particularly well-suited to well-structured domains (mathematics, reading, science procedures) where the content can be broken into small, sequenceable steps and where student mastery can be assessed through clear performance criteria.

Limitations. The principles are weighted heavily toward explicit, teacher-directed instruction. They say little about when to transition from guided to independent practice, how to develop inquiry skills, or how to handle ill-structured content. They also say little about motivation — the principles describe an efficient teaching machine, but they do not address the autonomy-support considerations that L1-002 identified as essential for sustained engagement. A classroom that rigidly follows all 10 principles may be effective for short-term learning but potentially damaging for long-term motivation if it leaves no room for student choice, exploration, or ownership.

Rosenshine's principles are best understood as a set of design guidelines for the explicit instruction phase of a broader instructional approach — what good direct instruction looks like. They are not a complete instructional theory.

7.2 MERRILL'S FIRST PRINCIPLES OF INSTRUCTION

David Merrill (2002) proposed five "First Principles" — features found across effective instructional programs regardless of theoretical orientation:

1. **Problem-centered:** Learning is promoted when learners are engaged in solving real-world problems.
2. **Activation:** Learning is promoted when existing knowledge is activated as a foundation for new knowledge.
3. **Demonstration:** Learning is promoted when new knowledge is demonstrated to the learner.
4. **Application:** Learning is promoted when new knowledge is applied by the learner.
5. **Integration:** Learning is promoted when new knowledge is integrated into the learner's world.

Merrill's framework is notable for its deliberate eclecticism — it claims to identify principles that are common to effective instruction regardless of whether the instruction is characterized as direct, constructivist, or inquiry-based. The framework is fundamentally a learning cycle: engage with a problem (activation), see how it works (demonstration), try it yourself (application), make it yours (integration).

Strengths. The framework bridges the instruction-inquiry divide by placing both demonstration (explicit instruction) and problem-solving (application) within a single cycle. The problem-centered principle is consistent with PBL and productive failure research, while the demonstration principle is consistent with the worked example effect. The activation principle aligns with Kapur's argument that activating prior knowledge before instruction enhances learning.

Limitations. The framework operates at a high level of abstraction. It tells you what features to include but not how to implement them in specific contexts. 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. And like Rosenshine, Merrill says relatively little about how to calibrate the balance of principles to learner expertise or domain characteristics.

7.3 THE 4C/ID MODEL: COMPLEX LEARNING MADE SYSTEMATIC

Van Merriënboer and Kirschner's Four-Component Instructional Design (4C/ID) model is the most sophisticated attempt to provide a comprehensive framework for instructional design in complex domains (van Merriënboer, Kirschner & Kester, 2003; van Merriënboer & Kirschner, 2018). The model has four components:

1. **Learning tasks:** Whole, authentic tasks of increasing complexity that form the backbone of the curriculum. Unlike Rosenshine's "small steps," the 4C/ID model insists on whole-task practice from the beginning — simplified versions of the whole task, not fragments of it. Tasks are organized in "task classes" of increasing difficulty, and within each class, learners progress from high-support to low-support versions.

2. **Supportive information:** Explanations of mental models and cognitive strategies that help learners perform the non-routine aspects of tasks. This information is presented before or during tasks and supports schema construction for the aspects of performance that require problem-solving and reasoning.

3. **Procedural information:** Just-in-time, step-by-step instructions for the routine aspects of tasks. Unlike supportive information, procedural information is presented exactly when needed and faded as the learner automates the procedures.

4. **Part-task practice:** Additional practice for sub-skills that need to be automatized to a high level of fluency. While the model insists on whole-task practice as the primary vehicle for learning, it acknowledges that some component skills (e.g., arithmetic operations in engineering, typing in computer programming) need dedicated practice to reach automaticity.

Strengths. The 4C/ID model is the framework that takes CLT most seriously as a design tool. It explicitly addresses the challenge of teaching complex, real-world skills that involve both routine and non-routine components. The distinction between supportive information (for schema construction) and procedural information (for automation) maps directly onto CLT's understanding of different types of learning. The insistence on whole-task practice from the beginning — progressively simplified, but always whole — addresses the transfer problem that arises when skills are taught in isolation.

The model also handles the expertise question well. The progression from high-support to low-support learning tasks within each task class, and the progression from simpler to more complex task classes, builds in the kind of scaffolding fading that the expertise reversal literature recommends. And the separate treatment of routine and non-routine aspects of tasks allows for different instructional approaches to each — explicit procedural instruction for routines, problem-based approaches for non-routines.

Limitations. The 4C/ID model is complex and requires substantial expertise to implement. The design process is resource-intensive — analyzing a complex skill into its component parts, designing authentic whole tasks at multiple complexity levels, creating supportive and procedural information for each, and building in part-task practice requires significant design capacity. The model has been implemented primarily in professional education (medicine, engineering, teacher education) and has limited evidence of implementation in K-12 settings.

The model also says relatively little about the motivational dimension. The emphasis on authentic tasks may serve motivation by providing relevance and meaning, but the framework does not explicitly address how to maintain autonomy support, manage the emotional challenges of complex task performance, or integrate assessment in motivation-preserving ways.

7.4 CHI'S ICAP FRAMEWORK

Michelene Chi's Active-Constructive-Interactive (ACI) framework (2009), later expanded to the ICAP framework with Ruth Wylie (Chi & Wylie, 2014), provides a taxonomy for classifying learning activities by their cognitive engagement level:

- **Passive:** Receiving information without overt processing — listening to a lecture, watching a video.
- **Active:** Manipulating information — highlighting, copying, paraphrasing, taking selective notes.
- **Constructive:** Generating new information beyond what was provided — self-explaining, creating analogies, drawing concept maps, generating hypotheses.
- **Interactive:** Collaborating to construct new understanding — building on each other's ideas, engaging in substantive dialogue, co-constructing explanations.

The framework predicts — and evidence supports — a hierarchy: Interactive > Constructive > Active > Passive for learning outcomes. The distinctions are precise and useful. Highlighting is "active" but produces little learning because it does not generate new knowledge structures. Self-explanation is "constructive" because it requires the learner to generate inferences not present in the material. Peer discussion is "interactive" when — and only when — both partners are generating and building on each other's ideas, not when one explains and the other listens.

Strengths. ICAP provides clear criteria for evaluating the cognitive depth of any learning activity, regardless of whether it is classified as "direct instruction" or "inquiry." A lecture in which students generate self-explanations (constructive) is more effective than a lecture in which students take verbatim notes (active), which is more effective than a lecture in which students just listen (passive). This framework cuts across the direct instruction-inquiry debate by showing that the critical variable is not the source of the activity (teacher-directed or student-directed) but the cognitive processing it demands.

The 2014 paper, with 2,417 citations and an FWCI of 60.68, has become one of the most influential frameworks in the field precisely because it provides a common language for evaluating learning activities that is independent of the direct-instruction-versus-inquiry framing.

Limitations. The framework says less about when each level is appropriate — the hierarchy suggests that interactive activities are always superior, but this ignores the expertise dimension. A constructive self-explanation task may be more effective for novices than an interactive discussion, because the discussion may impose additional cognitive load from social processing that novices cannot afford. The framework also does not address the question of what counts as "interactive" in sufficient detail — not all peer discussion involves genuine co-construction, and the conditions under which group interaction produces learning (rather than social loafing or pooling of ignorance) are not fully specified.

7.5 ENGELMANN'S DIRECT INSTRUCTION: THE EVIDENCE-BASED OUTCAST

No treatment of instructional design frameworks is complete without addressing Engelmann's Direct Instruction (DI) — note the capital letters, which distinguish Engelmann's specific program from the generic concept of direct instruction (lower case). Siegfried Engelmann developed DI in the 1960s as a highly structured, scripted instructional approach in which teachers follow carefully designed sequences, use precise language, incorporate frequent student responses, and provide immediate corrective feedback.

The evidence base for Engelmann's DI is remarkable. Project Follow Through, the largest educational experiment in U.S. history, compared multiple instructional approaches across over 200,000 students in the 1970s. DI produced the best outcomes — not just in basic skills (where one might expect a structured approach to excel) but also in cognitive problem-solving and self-esteem measures. Subsequent research has generally confirmed these findings. Stockard, Wood, Coughlin, and Rasplica Khoury (2018) conducted a meta-analysis of 328 studies spanning 50 years and found that Engelmann's DI produced positive outcomes across all student populations and all outcome measures, with effects that were "consistently positive" and larger than those found for most other educational interventions.

Yet Engelmann's DI remains one of the most unpopular approaches in education. It is criticized as overly rigid, dehumanizing, incompatible with constructivist philosophy, and — in a criticism with racial overtones — particularly applied to disadvantaged populations as a form of educational control. Teachers resist scripted instruction because it removes their professional judgment. Edu-

cation schools rarely teach it. Policy makers who advocate for it are accused of reducing education to compliance.

The disconnect between DI's evidence base and its adoption is one of the most revealing phenomena in education. Several explanations are plausible:

First, the philosophical objection is genuine, not merely aesthetic. Engelmann's DI is designed for maximum efficiency of knowledge transmission. It is not designed to develop inquiry skills, creative problem-solving, or the kind of independent thinking that many educators consider central to education's purpose. The evidence that DI produces good outcomes on standardized assessments does not address the question of whether it produces the kinds of learners that a democratic society needs. Kuhn's objection — that effectiveness depends on what you are trying to achieve — applies here with full force.

Second, the autonomy concern is real. L1-002 established that teacher autonomy support is essential for student motivation. A scripted curriculum in which the teacher reads from a script and students respond in unison offers minimal autonomy — for teachers or students. The motivational costs may not appear on short-term assessments but may accumulate over time, contributing to the kind of motivational decline across schooling that L1-002 documented. This is speculation, not established fact — the longitudinal motivational effects of Engelmann's DI have not been studied. But it is a legitimate concern.

Third, the scope of DI is limited. Engelmann's approach works best for well-structured content that can be broken into small, sequenceable steps with clear criteria for mastery. It is less applicable to ill-structured domains — writing, ethical reasoning, creative problem-solving, design thinking — where the content cannot be so cleanly decomposed. For a curriculum that aims to develop complex competencies, DI is a useful tool for some components (building foundational knowledge and automated skills) but cannot be the whole approach.

The relationship between Engelmann's DI and Rosenshine's Principles deserves clarification, because the terms are often confused. Rosenshine described what effective teachers do — the principles are descriptive of observed practice in high-gain classrooms. Engelmann prescribed a specific curriculum design — the program is a scripted instructional system based on logical analysis of concepts and carefully sequenced examples and non-examples. Both are "direct instruction" in the generic sense, but they differ in important ways. Rosenshine's principles leave room for teacher judgment and adaptation; Engelmann's scripts do not. Rosenshine's principles are compatible with a wider range of curricula; Engelmann's DI is a specific curriculum.

The tension between DI's evidence base and its unpopularity illuminates a deeper issue in education: the gap between what works for measurable outcomes and what educators believe education should look like. This is not merely a political or aesthetic disagreement. It is a genuine values conflict about the purposes of education — whether efficiency of knowledge transmission is the primary goal, or whether the development of autonomous, self-directed, inquiry-capable learners is equally or more important. The evidence cannot resolve this conflict because it depends on what outcomes are valued. But the evidence can clarify the tradeoffs, and the honest assessment is:

Engelmann's DI is probably the most effective approach for rapidly building foundational knowledge and skills in well-structured domains, particularly for students who start with significant knowledge deficits. Its evidence base should not be ignored. But it should not be treated as a complete instructional model, and its implementation should be balanced with approaches that develop autonomy, inquiry capacity, and the motivation for sustained learning. The question for curriculum design is not "DI or not?" but "DI for what, and alongside what else?"

8 THE HARD QUESTIONS

The preceding sections have established a coherent framework: the evidence supports an expertise-adaptive approach to instruction in which explicit instruction predominates for novices and inquiry-based approaches become increasingly appropriate as learners develop. But this framework has significant gaps and unresolved tensions. Acknowledging these honestly is essential.

8.1 THE ILL-STRUCTURED DOMAIN PROBLEM

The strongest evidence for the instructional design principles reviewed above — CLT, worked examples, the expertise reversal effect, productive failure — comes from well-structured domains, particularly mathematics, physics, and computer science. In these domains, problems have clear structure, solutions can be decomposed into steps, and correctness can be unambiguously assessed.

But much of what education aims to teach does not live in well-structured domains. Writing a persuasive essay, making an ethical judgment, designing a product, evaluating a historical argument, creating a work of art, developing a business strategy — these are ill-structured tasks where the problem space is not clearly defined, multiple solutions may be equally valid, and "correctness" is a matter of judgment, not calculation.

What does cognitive load theory have to say about teaching in ill-structured domains? Surprisingly little. The Lo survey identified this as Gap 5 — one of the most significant blind spots in the field — and this investigation confirms that the gap remains. Some efforts have been made to extend CLT to ill-structured domains. Kalyuga and Singh (2016) argued for "rethinking the boundaries of cognitive load theory in complex learning," proposing that the theory's principles apply but require different implementation strategies when element interactivity is high and problem structure is fluid. Van Merriënboer's 4C/ID model is the most serious attempt to handle complex, authentic tasks, but its evidence base in genuinely ill-structured domains (as opposed to complex but ultimately well-structured professional tasks) is limited.

What does the evidence suggest? Several tentative observations:

Worked examples look different in ill-structured domains. You cannot provide a worked example of "how to write a persuasive essay" in the same way you can provide a worked example of "how to solve a quadratic equation." But you can provide exemplars — high-quality examples of the target performance — accompanied by expert commentary that makes the reasoning behind design choices visible. The writing pedagogy literature suggests that studying exemplars with guided analysis is effective for developing writing skill (Graham & Perin, 2007). This is, functionally, a worked example adapted for an ill-structured domain.

Productive failure may be particularly powerful in ill-structured domains. If the mechanism of productive failure is activating prior knowledge, creating awareness of knowledge gaps, and preparing learners for subsequent instruction through comparison and contrast, these mechanisms may be even more valuable in domains where the problem space is not well-defined. Attempting to solve an ill-structured problem before instruction forces learners to confront the complexity of the domain in ways that studying a simplified model cannot. The Kapur research has been extended to domains beyond mathematics — science, design, and business — with generally positive results, though the evidence base is smaller.

The autonomy-structure balance may tilt differently. In ill-structured domains, the "correct answer" is not a fixed target but a space of acceptable solutions. This creates more natural room for learner autonomy — the learner can pursue their own approach within the problem space without the risk of "getting it wrong" in the way that a mathematics student can. This suggests that the transition from explicit instruction to guided inquiry can happen earlier in ill-structured domains, because the domain itself provides more tolerance for variation.

Feedback is harder but not impossible. L1-003 identified feedback for ill-structured tasks as Gap 2. Sadler's (2010) concept of "guild knowledge" — the tacit understanding of quality that experts develop through extensive experience — suggests that the assessment challenge in ill-structured domains is not that quality cannot be recognized, but that the criteria are difficult to articulate and communicate. The implication for instructional design is that ill-structured domains require more modeling of expert judgment — showing learners how an expert evaluates work, what criteria they attend to, how they weigh competing considerations — and more opportunities for guided practice in evaluation before students are expected to produce high-quality work independently.

The 4C/ID model provides the most promising theoretical bridge to ill-structured domains, because it was designed for complex, authentic tasks that combine routine and non-routine components. The model's insistence on whole-task practice (rather than decomposition into isolated sub-skills), its distinction between supportive information (for non-routine aspects) and procedural information (for routine aspects), and its emphasis on task classes of increasing complexity are all relevant to ill-structured domains. But the model's evidence base in genuinely ill-structured tasks (as opposed to complex but ultimately well-structured professional procedures like medical diagnosis or engineering design) remains limited.

The ICAP framework may also be particularly valuable for ill-structured domains. In domains where the "correct answer" is not clear-cut, the push toward constructive and interactive engagement becomes especially important. A student who passively receives a model essay learns little about writing. A student who actively annotates the essay learns something about its features. A student who constructively generates an analysis of why the essay works (identifying the argument structure, the evidence deployment, the rhetorical strategies) learns substantially more. And a student who interactively discusses the essay with peers, building on each other's analyses, learns the most — because the discussion reveals aspects of quality that no individual analysis captures. This is the ICAP hierarchy operating in an ill-structured domain, and it suggests that Chi's framework may be more broadly applicable than CLT's specific prescriptions.

The honest assessment is that the field of instructional design has a significant blind spot for ill-structured domains, and the most well-supported prescriptions (worked examples, completion problems, scaffolding fading) need adaptation — not abandonment — for domains where problem structure is fluid and quality is multidimensional. This is the most important gap identified in this investigation, and it has the most direct implications for Applied Pedagogy's curriculum design.

8.2 THE AUTONOMY-STRUCTURE TENSION

L1-002 identified the optimal autonomy-structure balance as Gap 3 — the central practical challenge of SDT-informed instruction. This investigation has made the tension sharper rather than resolving it.

The tension is this: CLT says novices need explicit instruction because unguided exploration overloads working memory. SDT says learners need autonomy support because controlling instruction undermines intrinsic motivation. Can both be true simultaneously?

They can, but the resolution requires distinguishing between the content of instruction and the context of instruction.

The content of instruction — what information is presented, in what sequence, using what format — is constrained by CLT. Novices need worked examples, small steps, and guided practice. This is not optional; it is determined by the architecture of human cognition.

The context of instruction — how the information is framed, what choices the learner has, what rationales are provided, what language is used — is where SDT operates. Explicit instruction can be delivered in an autonomy-supportive way or a controlling way. The difference lies not in the instructional technique but in the relational context.

Consider two versions of the same lesson:

Controlling version: "Today we will learn how to solve quadratic equations. Open your textbooks to page 47. Watch while I demonstrate. Now do problems 1–20. You must finish by the end of class."

Autonomy-supportive version: "Quadratic equations show up in all kinds of real situations — engineering, physics, even sports analytics. I want to show you a powerful technique for solving them, because once you have this tool, you'll be able to tackle problems that are currently out of reach. I'll demonstrate with a couple of examples, and then you'll choose from several practice sets depending on what feels like the right challenge level. Take a moment to look at the options."

Both lessons deliver the same content — worked examples followed by guided practice. The cognitive load is identical. But the motivational context is different. The second version provides rationale (supporting internalization), offers meaningful choice within structure (supporting autonomy), and frames the skill as a tool for the learner's own purposes (supporting competence). This is what Ahmadi et al. (2023) call "autonomy-supportive structure" — a concept that L1-002 highlighted as essential.

The practical implication is that the CLT prescription for explicit instruction and the SDT prescription for autonomy support are not in conflict — they operate on different dimensions of the instructional experience. You can give a clear, structured, worked-example-based lesson that also provides meaningful choices, genuine rationales, invitational language, and respect for the learner's perspective. You can also give a "student-directed inquiry" lesson that is actually controlling — where the illusion of choice masks a predetermined path and the teacher's disapproval of "wrong" approaches is palpable.

This resolution is real but partial. It handles the case where the content is well-structured and the prescription is clear. The harder case is when the learner needs to develop the ability to manage their own learning — the self-regulation capacity that L1-002 identified as a teachable meta-skill. Paradoxically, self-regulation must initially be taught through explicit instruction (because novice self-regulators do not know what strategies to use), but the goal is to develop the learner's capacity for autonomous, self-directed learning. The instruction must progressively hand over regulatory responsibility as the learner develops — a fading process analogous to the fading of worked examples, but operating at the metacognitive level.

The practical implications of this dual fading — content scaffolding and regulatory scaffolding — are significant. Consider a curriculum unit on scientific investigation. The content fading follows the expertise continuum: initially, students work through structured investigations with step-by-step guidance (worked examples of investigation); then they complete partially specified investigations (completion problems); then they design their own investigations with guidance; finally, they conduct independent investigations. The regulatory fading runs in parallel but is not identical: initially, the teacher models self-regulation explicitly ("I'm going to check my understanding by summarizing what I've learned so far — watch how I do this"); then the teacher prompts self-regulation at key moments ("What should you do before you move to the next step?"); then the

teacher provides self-regulation tools (checklists, rubrics) without prompting; finally, the students regulate their own learning independently.

The productive failure research adds another layer. Kapur's work shows that the emotional experience of struggle — when properly managed — can itself support motivation. The curiosity generated by an unresolved problem, the Zeigarnik effect that keeps the problem active in mind, the satisfaction of seeing one's failed attempts integrated into a complete understanding — these are motivational mechanisms that operate through productive failure, not despite it. The key is the "properly managed" qualification. Productive failure in a psychologically unsafe environment — where failure is punished, judged, or visible to hostile peers — produces anxiety and avoidance, not curiosity and engagement. The social surround that Kapur (2024) describes — re-norming expectations so exploration is valued, modeling vulnerability, creating psychological safety — is not a nice-to-have addition to the cognitive design. It is a necessary condition for the cognitive mechanisms to operate.

This convergence — CLT constraining the what, SDT informing the how, and productive failure showing that well-designed struggle can serve both cognitive and motivational goals simultaneously — represents the most promising direction for instructional design theory. But it is a demanding integration. It requires teachers who understand cognitive load well enough to design effective worked examples, who understand motivation well enough to deliver instruction autonomy-supportively, who understand productive failure well enough to design effective struggle tasks, and who have the diagnostic skill to know when to deploy each approach. This is expert teaching, and the field has not yet solved the problem of developing this expertise at scale.

8.3 CULTURAL VARIATION

Does the optimal instructional approach vary across cultures? The evidence is limited but suggestive. Japanese lesson study, in which teachers collaboratively design, observe, and refine lessons, produces instruction that often looks different from Western direct instruction — with more emphasis on whole-class problem-solving, student explanation, and multiple solution methods. French didactique, with its emphasis on carefully designed "situations" that create productive cognitive conflict, shares features with both productive failure and CLT. Scandinavian pedagogical traditions emphasize student autonomy and democratic participation in ways that go beyond SDT's prescriptions.

The CLT foundations should be universal — working memory limitations are features of human cognitive architecture, not cultural products. But the instructional implementations of those foundations may be culturally mediated. What counts as "scaffolding," what kinds of teacher-student interaction are effective, how feedback is received and processed, and what level of autonomy is experienced as supportive versus abandoning may all vary across cultural contexts.

There is also a deeper philosophical question about whether the goals of instruction vary across cultures in ways that affect the optimal approach. Western education tends to prioritize individual understanding and individual performance. East Asian education traditions place greater emphasis on collective practice and mastery through repetition — and produce world-leading performance on international assessments. Whether this reflects cultural preferences for different instructional approaches or different definitions of what counts as success is an open question. The Japanese mathematics classroom, for instance, typically involves students working on a single challenging problem for an extended period, followed by whole-class discussion of multiple solution methods — a structure that has elements of both productive failure and direct instruction, but organized differently from either Western model. The French tradition of didactique emphasizes carefully

engineered "didactical situations" that create cognitive conflict — a structured form of productive failure grounded in an entirely different theoretical tradition (Brousseau's theory of didactical situations) that arrived at similar conclusions from a different starting point.

The honest assessment is that almost all of the evidence reviewed in this dissertation comes from Western, educated, industrialized contexts — primarily the United States, the Netherlands, Australia, and Singapore. The cross-cultural evidence is thin, and the lab should flag this as a limitation whenever making instructional design recommendations. The convergence of Western CLT research and non-Western pedagogical traditions on similar principles (the value of structured struggle, the importance of expertise-adaptive instruction, the power of comparison and contrast) is suggestive that the underlying cognitive principles may be universal even if their implementation varies.

8.4 TRANSFER: THE UNSOLVED PROBLEM

Does instructional approach affect whether learning transfers to new contexts? This is arguably the most important question in education, and the answer remains frustratingly incomplete.

The meta-analytic evidence on PBL and inquiry learning consistently shows stronger effects on transfer and application outcomes than on immediate knowledge recall. Productive failure's primary advantage is in transfer, not in procedural competence. These findings suggest that more active, constructive, and generative learning approaches produce more transferable knowledge — knowledge that the learner can apply flexibly in new situations.

The mechanism is plausible. If transfer depends on having well-organized, deeply connected schemas (rather than isolated facts and procedures), then learning approaches that require the learner to generate connections, compare approaches, and construct understanding from multiple angles should produce more transferable knowledge than approaches that deliver information in pre-organized form. The ICAP framework's hierarchy — Interactive > Constructive > Active > Passive — predicts this pattern, and the evidence is broadly supportive.

But the evidence for far transfer — applying learned principles in genuinely novel, distant domains — remains elusive regardless of instructional approach. The Lo survey identified this as Gap 1, and nothing in this investigation resolves it. Near transfer (applying skills to similar tasks in similar contexts) is achievable with well-designed instruction. Far transfer (applying abstract principles across distant domains) remains the holy grail of education research — frequently promised, rarely delivered, and possibly overstated as a realistic educational goal.

Willingham (2009/2021) made this point accessible for practitioners: thinking is domain-specific because thinking requires content to think with. A student who is an excellent critical thinker in history (evaluating sources, identifying bias, constructing arguments from evidence) may be a poor critical thinker in science (designing experiments, evaluating statistical claims, distinguishing correlation from causation) — not because they lack some general "critical thinking skill," but because critical thinking in each domain requires domain-specific knowledge and domain-specific strategies. The Lo survey's conclusion that knowledge and skills are inseparable is directly relevant here: you cannot think critically without something to think critically about, and that something is domain knowledge.

The practical implication is that instructional design should be optimized for near transfer — designing instruction that helps learners apply what they learn to realistic variations of the tasks they practiced. This means using varied examples, multiple problem types, interleaved practice, and authentic contexts. It means teaching for flexible knowledge application, not just knowledge

acquisition. But it also means being honest about the limits of transfer and not promising that "teaching critical thinking" in one domain will produce critical thinkers across all domains.

There is a more hopeful reading of the transfer evidence, however, that should not be dismissed. While far transfer of specific skills or knowledge is rare, there is suggestive evidence that certain metacognitive dispositions — the habit of looking for analogies, the habit of checking one's understanding, the habit of seeking counterexamples — may transfer more readily because they operate at a higher level of abstraction. L1-002's finding that self-regulation can be taught explicitly suggests that at least the metacognitive component of transfer may be amenable to instruction. The question is whether metacognitive strategies, unlike domain-specific knowledge, can be taught in a way that transfers. The evidence is suggestive but not conclusive, and L1-002 noted that self-regulation tends to be domain-specific in practice even though the strategies are domain-general in principle.

Part IV

SYNTHESIS

9 SYNTHESIS

Drawing together the evidence from CLT, the instruction-inquiry debate, productive failure, the expertise reversal effect, and the practical frameworks, what does an evidence-based approach to instructional design look like?

9.1 THE CORE PRINCIPLES

1. Instruction should be expertise-adaptive. The single most important principle is that the optimal instructional approach depends on the learner's current expertise in the domain. What works for novices does not work for experts, and vice versa. This is not a matter of preference or philosophy — it is a consequence of human cognitive architecture.

2. For novices: explicit instruction is the foundation, but not the whole story. Novices need worked examples, small steps, clear explanations, and guided practice. But even novice instruction can benefit from brief preparatory activities — structured struggle, prediction, activation of prior knowledge — that prepare the learner to receive explicit instruction more effectively. And even novice instruction should be delivered in an autonomy-supportive context that provides rationale, choice, and respect.

3. Instruction should progressively fade scaffolding. As learners develop, the evidence overwhelmingly supports a gradual reduction in guidance: from complete worked examples to completion problems to guided practice to independent problem-solving. The transition should be based on demonstrated competence, not on a fixed schedule. Diagnostic assessment of learner expertise — informal or formal — should guide the fading process.

4. The outcome determines the approach. If the goal is factual knowledge or procedural competence, direct instruction is efficient and appropriate. If the goal is conceptual understanding or transfer, guided inquiry and productive failure have the advantage. If the goal is the development of inquiry skills themselves, inquiry practice is necessary. A well-designed curriculum pursues all of these goals and uses different approaches for each.

5. Assessment and feedback are integral to instruction, not separate from it. L1-003 established that retrieval practice is one of the most robust learning tools available, that feedback should be task-focused and process-focused, and that assessment should be frequent, low-stakes, and formative. These findings should be woven into the fabric of instruction. Every lesson should include retrieval practice from prior lessons. Every practice activity should generate feedback. Every assessment should inform the next instructional move.

6. The domain matters. Well-structured domains (mathematics, physics, programming) are well-served by the worked-example-to-problem-solving trajectory. Ill-structured domains (writing, ethics, design, entrepreneurship) require adapted approaches — exemplars with guided analysis, multiple solution paths, expert modeling of judgment, and greater tolerance for ambiguity. The 4C/ID model provides the most useful framework for complex domains that combine routine and non-routine components.

7. Motivation is not separate from instruction. The autonomy-structure tension is real but resolvable. Explicit instruction should be delivered in an autonomy-supportive context. Struggle should occur in psychologically safe environments. Practice should be separated from evaluation.

Choice should be offered within structure. Rationales should be genuine. These are not "nice-to-haves" — they are essential for sustained learning, as L1-002 established.

9.2 THE INSTRUCTIONAL DESIGN SEQUENCE

For any unit of curriculum, the evidence suggests a sequence:

1. Prepare. Activate prior knowledge. Create awareness of the knowledge gap that the unit will address. This can be as simple as a brief retrieval practice quiz on prerequisite knowledge, or as elaborate as a productive failure task in which learners attempt a problem they cannot yet solve. The preparation phase creates cognitive hooks for the instruction that follows.

2. Present. Deliver explicit instruction on the target knowledge and skills. Use worked examples for procedures. Use models and explanations for concepts. Use multimedia principles (coherence, signaling, contiguity, segmenting, pre-training) to minimize extraneous load. This is the phase where Rosenshine's principles apply most directly.

3. Practice. Provide structured practice with immediate feedback. Begin with high scaffolding (completion problems, guided practice) and gradually fade to independent practice. Interleave practice with review of prior topics. Keep practice low-stakes and separate from evaluation. Provide feedback that is task-focused and process-focused, never person-focused.

4. Apply. Present problems that require the learner to apply knowledge in new contexts. These should be varied, authentic, and increasingly complex. This is the phase where guided inquiry becomes appropriate — learners have sufficient knowledge to explore a problem space, and the exploration develops the flexible, transferable understanding that pure practice may not.

5. Reflect. Support metacognitive reflection — what was learned, how it was learned, what strategies were effective, what remains unclear. This develops the self-regulation capacity that L1-002 identified as essential for sustained learning. Reflection can be brief (a one-minute exit ticket) or extended (a structured self-assessment against criteria), but it should be explicit and regular.

6. Assess formatively. Use the results of practice, application, and reflection to assess where each learner stands and to inform the next instructional decision. If mastery has been achieved, move to more complex material or reduce scaffolding. If gaps remain, provide additional targeted instruction. This closes the formative assessment loop that L1-003 identified as one of the most powerful tools available.

This sequence is not rigid. The phases may overlap, repeat, or be reordered depending on context. The preparation phase may be extended into a full productive failure lesson for some topics. The presentation phase may be brief or absent when learners have sufficient knowledge to learn through guided inquiry. The application phase may be the primary vehicle for learning in ill-structured domains. The framework is descriptive of a general logic, not prescriptive of a fixed routine.

9.3 WHAT EACH FRAMEWORK CONTRIBUTES

The frameworks reviewed in this investigation each contribute something to the composite picture:

- **Rosenshine** tells us what good explicit instruction looks like — small steps, many questions, high success rates, guided practice, review.
- **Merrill** tells us what a complete learning experience contains — problem-centering, activation, demonstration, application, integration.
- **4C/ID** tells us how to design for complex, authentic performance — whole tasks, supportive and procedural information, part-task practice.

- **ICAP** tells us how to evaluate the cognitive depth of any learning activity — and that we should push toward constructive and interactive engagement whenever possible.
- **Engelmann** shows us what maximally effective explicit instruction looks like when efficiency is the primary goal — and reminds us that the evidence for structured instruction is stronger than most educators acknowledge.
- **Kapur** shows us that the conventional sequence (instruct → practice) is not always optimal — and that carefully designed struggle before instruction can deepen understanding and enhance transfer.

No single framework is complete. The curriculum designer needs all of them, deployed at appropriate moments and calibrated to the learner's developing expertise.

9.4 EVALUATION AGAINST THE COMPETENCE STACK

Applied Pedagogy defines competence as a five-layer stack: domain knowledge, skill, judgment, metacognition, and character/disposition (see COMPETENCE-TARGET.md). A critical evaluation of the instructional approaches reviewed in this investigation against this stack reveals an important asymmetry.

Most of the approaches reviewed are strongest at layers 1 and 2 — domain knowledge and skill. Worked examples build domain knowledge. Deliberate practice develops skill. Engelmann's DI excels at both. Rosenshine's Principles are designed for efficient knowledge and skill acquisition. Even the expertise reversal effect is fundamentally about optimizing the rate at which knowledge and skill develop.

Layer 3 — judgment — is where the picture begins to differentiate. Judgment requires exposure to varied, consequential, and ambiguous situations. This is precisely what well-designed PBL, guided inquiry, and productive failure provide that direct instruction alone does not. When students must decide which approach to take, evaluate whether a solution is adequate, or navigate a problem with no clean answer, they are practicing judgment. Merrill's First Principles (problem-centered) and the 4C/ID model (authentic whole tasks) are explicitly designed to develop judgment through practice with realistic complexity. Direct instruction and worked examples, by contrast, typically present pre-structured problems where the judgment calls have already been made.

Layer 4 — metacognition — is addressed by several of the approaches, though often implicitly rather than by design. Productive failure develops metacognitive awareness through the awareness mechanism (the "A" in Kapur's 4A model): struggling and failing makes learners conscious of what they do not know. The ICAP framework's constructive activities (self-explanation, generating hypotheses) require metacognitive processing. Formative assessment, as L1-003 documented, supports metacognition by providing information that learners can use to monitor and adjust their own learning. But few of the frameworks make metacognition an explicit design target. This is a significant gap.

Layer 5 — character and epistemic disposition — is the most neglected by every framework reviewed. Intellectual honesty, tolerance for uncertainty, and willingness to say "I don't know" are not targets of any standard instructional design model. The closest approach is Kapur's emphasis on the social surround — creating environments that normalize failure, value exploration, and reward honest engagement over performance. The productive failure approach, when properly implemented, cultivates exactly the dispositions Layer 5 describes: comfort with not-knowing, persistence through difficulty, willingness to be wrong publicly. But this is a byproduct of the design, not a primary target.

The implication for Applied Pedagogy is clear: no existing instructional design framework adequately addresses all five layers of the competence stack. Layers 1–2 are well-served by the evidence base reviewed here. Layer 3 is partially addressed by inquiry-based and problem-centered approaches. Layers 4–5 require deliberate design that goes beyond what any single framework provides. This is where Applied Pedagogy's curriculum must go further than the existing frameworks — integrating explicit metacognitive instruction (L1-002's finding that self-regulation can be taught), environmental design that rewards honesty and tolerates error (the social surround), and assessment practices that treat error as information rather than failure (L1-003's findings). The instructional design evidence tells us how to build layers 1–3. Layers 4–5 require a synthesis that the field has not yet achieved.

10 THE CONNECTION TO MOTIVATION AND ASSESSMENT

This investigation does not stand alone. It builds on — and must be integrated with — the findings from the motivation investigation (L1-002) and the assessment investigation (L1-003).

10.1 THE MOTIVATION CONNECTION

L1-002 established that:

- Autonomy support, competence support, and relatedness support (SDT's three basic needs) are essential for sustained motivation.
- Self-regulation can and should be taught explicitly, within domains.
- Extrinsic rewards reliably undermine intrinsic motivation.
- Productive failure is a high-skill teaching practice that requires careful motivational management.

These findings constrain instructional design in important ways. An instructional approach that is cognitively optimal but motivationally destructive will fail in the long run. Specifically:

Explicit instruction must be delivered autonomy-supportively. The CLT prescription for worked examples and guided practice is not a license for controlling, compliance-oriented instruction. The language matters. The choice architecture matters. The rationale matters.

The transition to greater independence must be genuine, not cosmetic. As scaffolding fades, the learner must experience real ownership of their learning — making genuine decisions, pursuing genuine interests, managing genuine challenges. A fake version of autonomy — where the learner chooses between two equally predetermined paths — is experienced as controlling and undermines motivation.

Struggle must be safe. The productive failure insight — that struggling before instruction deepens learning — requires a psychological safe zone in which failure is expected, normalized, and informational. If struggle is evaluated, judged, or punished, its cognitive benefits are overwhelmed by its motivational costs. This is why L1-002 and L1-003 both insist on separating practice from evaluation.

10.2 THE ASSESSMENT CONNECTION

L1-003 established that:

- Retrieval practice (the testing effect) is one of the most robust tools for learning.
- Feedback should be task-focused, process-focused, and actionable.
- Grades negate the benefit of feedback.
- Assessment should be frequent, low-stakes, and formative.

These findings are not separate from instructional design — they are part of it. Assessment is not something that happens after instruction; it is an integral component of instruction. Every lesson should begin with retrieval practice (leveraging the testing effect and the spacing effect). Every

practice activity should generate immediate, task-focused feedback. Every formative assessment should inform the next instructional decision.

The integration looks like this: the "Prepare" phase of the instructional sequence should include retrieval practice on prior learning — 3-5 minutes of low-stakes recall that strengthens prior learning while activating it for the current lesson. The "Practice" phase should include immediate corrective feedback on every attempt — not grades, but information about what was done correctly and what needs adjustment. The "Assess formatively" phase should generate data that drives the next instructional decision — whether to reduce scaffolding, provide additional practice, or move to new material.

This integration of instruction, assessment, and motivation is the practical expression of what the Lo survey called the most important underresearched area in the field: the assessment-motivation intersection. The evidence points toward an integrated design in which instruction, assessment, and motivation support are woven together — not treated as separate concerns managed by separate systems.

11 CLOSING ASSESSMENT

11.1 WHAT THE EVIDENCE CLEARLY SUPPORTS (HIGH CONFIDENCE)

The expertise reversal effect is real and consequential. Instruction that is optimal for novices becomes suboptimal or harmful as expertise develops. This is one of the most robust findings in instructional design research, replicated across multiple domains and research groups. *Confidence: High.*

Unguided discovery is ineffective for novices. Purely unstructured exploration, without scaffolding or feedback, does not produce effective learning for learners who lack relevant prior knowledge. The cognitive load argument is sound and empirically supported. *Confidence: High.*

Guided inquiry is effective when properly scaffolded. Multiple meta-analyses converge on this finding. The Lazonder and Harmsen (2016) meta-analysis (FWCI 209.59) showed substantial effects of guided inquiry on learning outcomes. *Confidence: High.*

The worked example effect is robust for novices in well-structured domains. Decades of research across multiple domains confirm that novices learn more efficiently from studying worked examples than from attempting to solve equivalent problems. *Confidence: High.*

Productive failure reliably enhances conceptual understanding and transfer relative to direct instruction. The meta-analytic evidence across 50+ studies is consistent and large. The mechanism (cognitive preparation for subsequent instruction) is theoretically coherent and empirically supported. *Confidence: High.*

The outcome measure determines which approach "wins." Direct instruction tends to win on immediate knowledge tests. Inquiry and productive failure tend to win on conceptual understanding, transfer, and long-term retention. This is not a minor methodological footnote — it is central to interpreting the evidence honestly. *Confidence: High.*

ICAP's hierarchy (Interactive > Constructive > Active > Passive) is broadly supported. The framework provides useful criteria for evaluating the cognitive depth of learning activities, and the evidence generally supports the predicted hierarchy. *Confidence: High, with the caveat that the hierarchy may interact with expertise level.*

11.2 WHAT THE EVIDENCE SUPPORTS WITH IMPORTANT CAVEATS (MEDIUM CONFIDENCE)

The expertise-adaptive developmental trajectory is the right framework, but the implementation details are thin. The principle that instruction should change as learners develop is well-supported. How to assess where a learner is on the expertise continuum, how to time the transitions, and how to manage this in heterogeneous classrooms — these practical questions lack well-tested answers. *Confidence: Medium.*

The autonomy-structure tension is resolvable in principle but challenging in practice. The distinction between the content of instruction (constrained by CLT) and the context of instruction (where SDT operates) is theoretically sound. Whether teachers can reliably implement autonomy-supportive explicit instruction at scale, especially under time pressure and with large classes, is an open empirical question. *Confidence: Medium.*

The 4C/ID model is the most promising framework for complex learning, but its evidence in K-12 settings is limited. The model is theoretically sophisticated and well-grounded in CLT, but most of its implementation evidence comes from professional education. Whether it can be adapted to the constraints of K-12 schooling is uncertain. *Confidence: Medium.*

Productive failure works beyond mathematics, but the evidence base is narrower. Extensions to science, business, and other domains show promise, but the depth of evidence is much less than in mathematics. The design principles (accessible language, contextualized problems, multiple solution paths, contrasting cases) are domain-general in principle but domain-specific in implementation. *Confidence: Medium.*

11.3 WHAT REMAINS GENUINELY UNCERTAIN (LOW CONFIDENCE)

How to teach effectively in ill-structured domains. The cognitive science of instruction has a significant blind spot for domains like writing, ethics, design, and creative problem-solving. The principles from well-structured domains (worked examples, fading, scaffolding) can be adapted, but how to adapt them — and whether the adaptations preserve the principles' effectiveness — is largely untested. *Confidence: Low.*

Whether instructional approach affects far transfer. Near transfer is achievable with well-designed instruction. Far transfer remains elusive regardless of approach. Whether inquiry-based instruction produces more transferable knowledge than direct instruction in a meaningful, practically significant way — rather than just a statistically significant way — is genuinely uncertain. *Confidence: Low.*

Cultural variation in optimal instructional approaches. Almost all of the evidence comes from Western contexts. Whether the same principles apply in East Asian, African, or Latin American educational cultures — and whether the same implementations are appropriate — is unknown. *Confidence: Low.*

Long-term motivational effects of different instructional approaches. Whether a curriculum based on Engelmann's DI produces motivational damage over years (as SDT would predict) or whether a curriculum based on productive failure produces motivational benefits over years (as the theory would predict) is unknown. The evidence is almost entirely short-term. *Confidence: Low.*

Engelmann's DI: effect sizes versus motivational costs. The evidence for DI's effectiveness on standardized outcomes is strong. The concern about its motivational effects is theoretically grounded but empirically untested. Whether DI's short-term achievement benefits are offset by long-term motivational costs remains an open question. *Confidence: Low regarding the long-term tradeoff.*

12

BEYOND THE BINARY

The direct instruction versus inquiry debate has consumed enormous intellectual energy for two decades. It has produced important insights — about the cognitive architecture constraints on learning, about the conditions under which different approaches are effective, about the role of scaffolding and guidance, about the importance of outcome measures. But the debate's binary framing has also been a distortion, obscuring the more nuanced picture that the evidence supports.

The evidence supports an expertise-adaptive model of instruction. For novices, explicit instruction with worked examples is the most efficient path to initial competence — but even novice instruction can be enhanced by preparatory activities that activate prior knowledge and create cognitive readiness. As learners develop, instruction should progressively shift toward guided inquiry, productive failure, and independent problem-solving, with the transitions calibrated to demonstrated expertise. The optimal approach also depends on the learning goal (knowledge, understanding, transfer, or inquiry capacity), the domain (well-structured versus ill-structured), and the motivational context (autonomy-supportive versus controlling).

This is not a wishy-washy "it depends" — it is a specific, evidence-grounded account of how instruction should change across the arc of learning. The arc looks like this:

Explicit instruction → faded scaffolding → productive failure → guided inquiry → independent investigation

Each stage builds on the previous one. Explicit instruction builds the foundational schemas that make productive failure possible. Productive failure builds the deep understanding that makes guided inquiry productive. Guided inquiry builds the inquiry skills that make independent investigation possible.

The practical implications for Applied Pedagogy's curriculum design are substantial and specific, and they are detailed in the practical-implications.md file that accompanies this dissertation. But the headline is this: the curriculum should not be organized around a single instructional philosophy. It should be organized around the learner's developmental trajectory in each domain, using different instructional approaches at different stages, integrated with the formative assessment practices that make the transitions visible and the autonomy-supportive practices that make sustained learning possible.

Consider what this means concretely for a curriculum unit on, say, statistics — a domain where the research evidence is particularly strong because so much of the productive failure work has been done there. The unit might begin with a productive failure task: students are given data on two basketball players and asked to determine who is more consistent. They cannot yet solve this correctly — they lack the concept of standard deviation — but they can generate measures of consistency (range, average deviation, visual comparisons) that capture parts of the idea. This struggle activates prior knowledge about variability, creates awareness of what their intuitive measures miss, generates emotional investment in the problem, and produces a set of student-generated solutions that the teacher can build on.

The instruction phase follows: the teacher takes the students' solutions, compares and contrasts them, shows what each captures and what each misses, and systematically builds toward the canonical concept of standard deviation. This is explicit instruction — clear, structured, step-by-step — but it is explicit instruction that connects to what students have already thought and felt

about the problem. The schemas being constructed are richer because they are connected to the students' own prior attempts.

Practice follows, with fading scaffolding: first, completion problems where the calculation is partially worked out; then, varied problems with immediate feedback; then, problems that require choosing the appropriate measure of variability for different situations (transfer). Assessment is woven throughout — retrieval practice at the start of each session drawing on prior concepts, formative checks during practice, and periodic low-stakes quizzes that test conceptual understanding as well as procedural competence.

As students develop competence with basic statistical concepts, the instruction shifts. They begin designing their own investigations — choosing data sets, selecting appropriate analyses, interpreting results. The teacher's role shifts from demonstrator to coach. The scaffolding is largely removed. And the motivation is sustained because students are doing genuine statistical work on questions they care about, not just executing procedures on textbook exercises.

This is the expertise-adaptive model in action. It is not exotic. It is not theoretically mysterious. It is the integration of what CLT, productive failure, formative assessment, and SDT all point toward: a developmental trajectory of instruction that starts with structured support and progressively releases responsibility as the learner develops the schemas and the self-regulation to handle it.

The defining challenge remains the implementation. The evidence tells us what to do; the practice of doing it — in real classrooms, with real teachers, for real learners at different stages — is the work that lies ahead.

Dissertation complete.

BIBLIOGRAPHY

Ahmadi, A., Noetel, M., Parker, P., Ryan, R. M., Ntoumanis, N., Reeve, J., Beauchamp, M., Dicke, T., Yeung, A., Ahmadi, M., Bartholomew, K., Chiu, T. K. F., Curran, T., Haerens, L., Kharazmi, Z. N., Lonsdale, C., & Marsh, H. (2023). A classification system for teachers' motivational behaviors recommended in self-determination theory interventions. *Journal of Educational Psychology*, 115(9), 1158–1176.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research*, 70(2), 181–214.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, 8(4), 293–332.

Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, 1(1), 73–105.

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

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.

de Jong, T. (2009). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science*, 38, 105–134.

de Jong, T., Lazonder, A. W., Chinn, C. A., Fischer, F., Gobert, J. D., Hmelo-Silver, C. E., Koedinger, K. R., Krajcik, J., Kyza, E. A., Linn, M. C., Pedaste, M., Scheiter, K., & Zacharia, Z. C. (2023). Let's talk evidence — The case for combining inquiry-based and direct instruction. *Educational Research Review*, 39, 100536.

Dochy, F., Segers, M., Van den Bossche, P., & Gijbels, D. (2003). Effects of problem-based learning: A meta-analysis. *Learning and Instruction*, 13(5), 533–568.

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

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

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

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.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, 19, 509–539.

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

Kalyuga, S., & Renkl, A. (2009). Expertise reversal effect and its instructional implications: Introduction to the special issue. *Instructional Science*, 38, 209–215.

Kalyuga, S., & Singh, A.-M. (2016). Rethinking the boundaries of cognitive load theory in complex learning. *Educational Psychology Review*, 28, 831–852.

Kalyuga, S., & Sweller, J. (2004). Measuring knowledge to optimize cognitive load factors during instruction. *Journal of Educational Psychology*, 96(3), 558–568.

Kapur, M. (2008). Productive failure. *Cognition and Instruction*, 26(3), 379–424.

Kapur, M. (2009). Productive failure in mathematical problem solving. *Instructional Science*, 38, 523–550.

Kapur, M. (2010). A further study of productive failure in mathematical problem solving: Unpacking the design components. *Instructional Science*, 40, 523–550.

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

Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *Journal of the Learning Sciences*, 21(1), 45–83.

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

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.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. A. (2012). The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, 36(5), 757–798.

Kuhn, D. (2007). Is direct instruction an answer to the right question? *Educational Psychologist*, 42(2), 109–113.

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

Loibl, K., Roll, I., & Rummel, N. (2017). Towards a theory of when and how problem solving followed by instruction supports learning. *Educational Psychology Review*, 29, 693–715.

Mayer, R. E. (2002). Multimedia learning. *Psychology of Learning and Motivation*, 41, 85–139.

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

Pedaste, M., Mäeots, M., Siiman, L. A., de Jong, T., van Riesen, S. A. N., Kamp, E. T., Manoli, C. C., Zacharia, Z. C., & Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review*, 14, 47–61.

Renkl, A., Atkinson, R. K., Maier, U. H., & Staley, R. (2002). From example study to problem solving: Smooth transitions help learning. *Journal of Experimental Education*, 70(4), 293–315.

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

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

Savery, J. R. (2006). Overview of problem-based learning: Definitions and distinctions. *Interdisciplinary Journal of Problem-Based Learning*, 1(1).

Schmidt, H. G., Loyens, S. M. M., van Gog, T., & Paas, F. (2007). Problem-based learning is compatible with human cognitive architecture: Commentary on Kirschner, Sweller, and Clark (2006). *Educational Psychologist*, 42(2), 91–97.

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

Sinha, T., & Kapur, M. (2021). Robust effects of the efficacy of explicit failure-driven scaffolding in problem-solving prior to instruction: A replication and extension. *Learning and Instruction*, 75, 101488.

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? A meta-synthesis of meta-analyses comparing PBL to conventional classrooms. *Interdisciplinary Journal of Problem-Based Learning*, 3(1).

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4(4), 295–312.

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, 2(1), 59–89.

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3), 251–296.

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

van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher-student interaction: A decade of research. *Educational Psychology Review*, 22, 271–296.

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

van Merriënboer, J. J. G., Kirschner, P. A., & Kester, L. (2003). Taking the load off a learner's mind: Instructional design for complex learning. *Educational Psychologist*, 38(1), 5–13.

Walker, A., & Leary, H. (2009). A problem based learning meta analysis: Differences across problem types, implementation types, disciplines, and assessment levels. *Interdisciplinary Journal of Problem-Based Learning*, 3(1).

Willingham, D. T. (2009/2021). *Why Don't Students Like School?* (2nd ed.). Jossey-Bass.

Zhang, L., Kirschner, P. A., Cobern, W. W., & Sweller, J. (2022). There is an evidence crisis in science educational policy. *Educational Psychology Review*, 34, 1157–1176.