THE COGNITIVE SCIENCE OF LEARNING

What We Know, Where It Breaks Down, and What It Means for Instruction

Applied Pedagogy Research Lab

Guido Bartolucci, Principal Investigator

guido@appliedpedagogy.com

LAB.APPLIEDPEDAGOGY.COM

L1-001 · March 2026

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

CONTENTS

Part I

FOUNDATIONS

1

THE STATE OF THE EVIDENCE

The cognitive science of learning has the strongest empirical foundation of any domain in the broader field of learning science. This is not a contested claim — it is the consensus of the field, reflected in meta-analyses, large-scale replications, and the judgment of researchers who work across traditions. When the Lo survey described the cognitive science evidence as "the most replicated in all of psychology," it was not overstating the case. The testing effect, the spacing effect, the worked example effect, and the basic architecture of working memory have survived decades of scrutiny.

But strength of evidence is not the same as completeness, and the distance between a laboratory demonstration and a practical curriculum is longer than the field's confidence sometimes suggests. This review takes the Lo survey's assessment as its starting point and asks the harder questions: What exactly do we know with confidence? Where do the well-established findings have boundary conditions that practitioners need to understand? Where does the evidence simply run out? And most importantly: what can a curriculum designer at Applied Pedagogy actually rely on from cognitive science — not as general principles, but as specific guidance for specific design decisions?

The answers that emerge are more nuanced than the headlines suggest. Cognitive load theory is powerful but has evolved significantly since its original formulation, and its central concept — germane load — has been effectively abandoned by its own creators. Retrieval practice works, but its effect sizes in authentic classrooms are smaller than laboratory studies predict. The spacing effect is one of the oldest findings in psychology, but translating optimal spacing intervals into curriculum schedules remains genuinely difficult. Productive failure produces impressive effects, but depends on design conditions that are easy to describe and hard to implement. And the question of transfer — whether learners can apply what they learn in one context to genuinely different contexts — remains the field's deepest unsolved problem.

The trajectory of the field since 2019 has been one of maturation rather than revolution. Cognitive load theory has been refined through the formalization of element interactivity as its central construct (Chen, Paas & Sweller, 2023) and expanded to address collaborative learning (Kirschner, Sweller, Kirschner & Zambrano, 2018) and digital environments (Skulmowski & Xu, 2021). Retrieval practice has moved from laboratory demonstrations to systematic classroom evidence (Agarwal, Nunes & Blunt, 2021). Productive failure has produced a book-length synthesis and meta-analytic evidence covering over fifty studies (Kapur, 2025). And the expertise reversal effect has received its first comprehensive meta-analysis (Tetzlaff, Simonsmeier, Peters & Brod, 2025), confirming its robustness while revealing important asymmetries.

What follows is an attempt to be honest about all of this — to give the practitioner a map that shows both the well-paved roads and the unmarked trails, and to be explicit about where the map simply says "here be dragons."

2 COGNITIVE LOAD THEORY: THE CENTRAL FRAMEWORK

2.1 THE ARCHITECTURE THAT CONSTRAINS EVERYTHING

Any serious discussion of the cognitive science of learning must begin with working memory, because working memory is the bottleneck through which all conscious learning must pass. Cowan (2001) established that working memory can hold approximately four chunks of novel information simultaneously — not seven, as Miller (1956) had famously suggested, but four. Long-term memory, by contrast, is effectively unlimited in both capacity and duration. The fundamental challenge of learning, from a cognitive science perspective, is moving information from the narrow bottleneck of working memory into the vast storage of long-term memory in organized, retrievable form.

Cognitive load theory (CLT), developed by John Sweller beginning in the 1980s and refined over four decades, is built on this architectural constraint. Its core insight is simple: if instructional materials require learners to simultaneously process more novel elements than working memory can handle, learning will fail — not because the material is intrinsically too difficult, but because the cognitive demands of processing it exceed the available capacity. The theory's power comes not from this general principle, which is almost tautological, but from its specific predictions about which instructional designs will overload working memory and which will manage it effectively.

2.2 THE THREE LOADS — AND WHY THERE ARE REALLY ONLY TWO

CLT originally distinguished three types of cognitive load. Intrinsic load is determined by the inherent complexity of the material and the learner's prior knowledge — learning to add single digits imposes lower intrinsic load than learning to solve differential equations. Extraneous load is imposed by poor instructional design — a diagram that requires learners to flip between pages to find explanatory text imposes extraneous load that a well-integrated diagram avoids. Germane load was defined as the productive cognitive effort directed at building schemas — the mental work of organizing new information into coherent knowledge structures.

The tripartite distinction was elegant in theory but problematic in practice. The difficulty was that germane load proved nearly impossible to distinguish from intrinsic load, either theoretically or empirically. If a learner is exerting cognitive effort to build a schema for a complex topic, is that effort germane (because it is directed at learning) or intrinsic (because it is determined by the material's complexity)? The distinction depended on the learner's intention, not on any observable property of the cognitive processing, which made it unmeasurable.

Sweller effectively resolved this problem in the 2019 update to the theory (Sweller, van Merriënboer & Paas, 2019) by collapsing germane load into intrinsic load. The revised position holds that there are really only two types of load: intrinsic (determined by element interactivity — more on this below) and extraneous (determined by instructional design). What was once called germane load is simply the intrinsic load of the material being processed productively, not a separate category. This simplification made the theory more parsimonious and more testable.

However, the issue is not fully resolved. Klepsch and Seufert (2020) developed and validated a questionnaire that measures all three types of cognitive load and found that the three-way distinction has prognostic validity — variations in difficulty are reflected in intrinsic load ratings,

variations in design are reflected in extraneous load ratings, and variations in deeper learning activities are reflected in germane load ratings. This suggests that learners experience something that feels like germane load, even if it may not be theoretically distinct from intrinsic load. The practical significance of this debate matters less than it might seem: the instructional design implications are the same regardless of whether germane load is a real category. Reduce extraneous load. Manage intrinsic load through sequencing and scaffolding. Direct the freed capacity toward productive processing.

2.3 ELEMENT INTERACTIVITY: THE CORE OF THE THEORY

If CLT has a single most important concept, it is element interactivity — and for most of the theory's history, this concept was underspecified. Chen, Paas, and Sweller (2023) addressed this gap in a paper that formalizes element interactivity as the theory's central construct for defining and measuring task complexity.

Element interactivity refers to the number of information elements that must be processed simultaneously for understanding to occur. A task with low element interactivity can be learned element by element — memorizing vocabulary words, for instance, where each word-meaning pair is largely independent of the others. A task with high element interactivity requires the learner to process multiple elements in relation to each other — understanding how supply and demand interact to determine market equilibrium, for instance, where understanding any single element requires understanding its connections to the others.

The critical insight is that element interactivity is not a fixed property of the material; it depends on the learner's prior knowledge. For an expert, what appears to be a complex multi-element task is actually a single chunk — a schema in long-term memory that can be activated as a unit, bypassing working memory limitations entirely. For a novice encountering the same material, each element must be processed individually, and the interactions between elements impose enormous demands on working memory. This is why the same instructional approach can be perfectly appropriate for one learner and completely wrong for another. It is also why the expertise reversal effect is not an anomaly but a direct prediction of the theory's core mechanism.

Chen, Paas, and Sweller (2023) argue that element interactivity can be estimated by counting the number of interacting information elements in a task, but they acknowledge that the knowledge held in long-term memory — which determines how many elements are truly novel — can only be estimated through teacher judgment or knowledge tests. This is an important limitation: the theory's central variable is not directly observable. It can be estimated, but the estimation requires knowing what the learner already knows, which introduces the very assessment challenges that make teaching difficult in the first place.

2.4 THE SPECIFIC EFFECTS: WHAT HAS HELD UP

CLT has generated a family of specific instructional effects, each derived from the theory's predictions about how instructional design affects cognitive load. Several of these effects are among the most well-replicated findings in educational research.

The worked example effect. Novice learners learn more effectively from studying worked examples than from solving equivalent problems. The explanation is straightforward: novice problem-solving imposes high extraneous cognitive load because the learner must search for solution methods — a process called means-ends analysis — which consumes working memory capacity that could otherwise be directed toward understanding the solution structure. Worked

examples eliminate this search cost by providing the solution, freeing cognitive capacity for schema construction. Barbieri, Miller-Cotto, Clerjuste, and Chawla (2023) conducted a meta-analysis of the worked example effect in mathematics and confirmed its robustness, finding that it generalizes across different types of mathematical content and learner populations. The effect is most pronounced for genuinely novel material where the learner lacks relevant schemas. This is a critical boundary condition: as expertise develops, the worked example effect diminishes and eventually reverses.

The split-attention effect. When learners must mentally integrate information from sources that are physically or temporally separated — a diagram on one page and its explanation on another, or a narrated animation where the narration is asynchronous with the visual — cognitive load increases and learning suffers. The solution is to physically integrate related information: labels directly on diagrams, narration synchronized with animation, text placed immediately adjacent to the visual elements it describes. This effect is the foundation of Mayer's spatial and temporal contiguity principles for multimedia learning and has been extensively replicated.

The redundancy effect. Presenting the same information in multiple forms — for example, reading aloud text that is simultaneously displayed on screen — can impair learning because the learner must process both presentations and reconcile them, even though they are redundant. This is counterintuitive: surely more is better? But the cognitive cost of processing redundant information is real, and eliminating genuinely redundant material reduces extraneous load. The boundary condition here is important: the effect applies to truly redundant information, not to complementary information presented in different modalities.

The modality effect. Presenting verbal and visual information in different modalities — spoken narration with diagrams, rather than written text with diagrams — takes advantage of the dual-channel nature of working memory (Baddeley's phonological loop and visuospatial sketchpad). By distributing the processing load across channels rather than concentrating it in a single channel, total effective working memory capacity is increased. This effect has been extensively replicated in multimedia learning contexts and is one of the most practically useful CLT findings for instructional designers working with video and interactive media.

2.5 THE EFFECTS UNDER PRESSURE: REPLICATION AND REVISION

Sweller (2023) addressed the replication crisis directly, arguing that CLT's relationship to replication failures has been productive rather than destructive. When empirical results contradicted CLT's predictions, the response was typically theory refinement rather than abandonment — and in each case, the refinement led to a more nuanced and accurate theory. The expertise reversal effect, for instance, emerged precisely because the worked example effect failed to replicate with expert learners. The redundancy effect was discovered because the modality effect failed when the verbal information was text rather than narration. In Sweller's account, CLT's "failures" are evidence of the theory's capacity for self-correction.

This is a generous reading, and it has some merit. CLT's core predictions about working memory limitations and the effects of extraneous cognitive load have survived replication challenges well. No credible line of research has overturned the basic finding that poor instructional design imposes cognitive costs, or that novices and experts require different instructional approaches.

But there is a less generous reading available too. De Jong (2009) raised concerns that CLT's predictions are sometimes ambiguous, particularly when intrinsic and extraneous load interact in ways the theory does not clearly specify. More broadly, CLT is a framework theory — its predictions are often qualitative rather than quantitative, and its central variable (element interactivity) is not

directly measurable. This makes it somewhat resistant to falsification: when a prediction fails, the theory can accommodate the failure by adding a new effect or refining a boundary condition, without ever being definitively wrong. This is not unique to CLT — most cognitive theories share this property — but it means that the theory's impressive track record should be understood partly as a reflection of its flexibility rather than purely as a reflection of its accuracy.

A retraction search for CLT-related work found no retractions of core CLT papers. This is noteworthy and reassuring. The field's empirical base appears sound.

2.6 EXTENSIONS: COLLABORATIVE AND DIGITAL LEARNING

Two significant extensions of CLT deserve attention because they address contexts where the original theory was silent.

Collaborative cognitive load theory. Kirschner, Sweller, Kirschner, and Zambrano (2018) extended CLT to collaborative learning, arguing that collaboration can be understood as a means of expanding effective working memory capacity. When a task has high element interactivity — too many elements for any individual working memory to handle — distributing the processing across multiple learners can reduce per-person cognitive load. But collaboration also introduces costs: the need to communicate, coordinate, and integrate understanding with other people imposes its own cognitive demands, which the authors term "transaction costs." Collaboration is beneficial when the task's element interactivity exceeds individual working memory capacity by enough to justify the transaction costs. For simple tasks with low element interactivity, collaboration is counterproductive — the transaction costs exceed the benefits of distributed processing.

This analysis yields a clear prediction: collaboration should be reserved for tasks that are genuinely too complex for individuals. Group work on simple tasks is not merely neutral but actively harmful, because it imposes coordination costs without reducing the cognitive load of the task itself. This has practical implications for curriculum design that run counter to the widespread assumption that collaboration is always beneficial.

Cognitive load in digital environments. Skulmowski and Xu (2021) reconceptualized extraneous cognitive load for digital and online learning, arguing that the traditional distinction between extraneous and intrinsic load becomes blurred when learners interact with complex digital interfaces. Navigating a learning management system, interpreting multimedia presentations, and managing multiple windows or tabs all impose cognitive demands that are not clearly categorizable as either extraneous (design-imposed) or intrinsic (material-imposed). They propose that the concept of extraneous load needs to be expanded to account for the "functional" demands of digital environments — demands that are not part of the content but are necessary for accessing the content.

Schneider, Beege, Nebel, Schnaubert, and Rey (2021) went further, proposing the Cognitive-Affective-Social Theory of Learning in digital Environments (CASTLE), which extends CLT by integrating affective and social factors that are particularly salient in digital learning contexts. CASTLE argues that cognitive load cannot be understood in isolation from the emotional and social demands of the learning environment — a position that brings CLT closer to the motivational and sociocultural perspectives that it has historically ignored.

Part II

THE EVIDENCE BASE

3 RETRIEVAL PRACTICE AND THE TESTING EFFECT

3.1 THE CORE FINDING

If cognitive load theory describes the architectural constraints on learning, retrieval practice is the single most effective strategy for working within those constraints. The testing effect — the finding that practicing retrieval produces better long-term retention than restudying — is among the most robust findings in cognitive psychology.

Karpicke and Roediger (2008), in a paper published in *Science*, demonstrated that students who practiced retrieving information from memory retained significantly more over a one-week delay than students who restudied the same material, even though the restudying students predicted they would perform better. The effect is not small: across numerous replications, the advantage of retrieval practice over restudying is both statistically significant and practically meaningful.

The mechanism is not merely additional exposure. Retrieving information from memory is itself a learning event — it strengthens the memory trace in ways that passive re-exposure does not. Roediger and Butler (2010) reviewed the evidence and concluded that the act of retrieval modifies the memory trace, making it more accessible for future retrieval. This is a genuinely surprising finding: the most effective way to study is not to re-encounter the material but to try to produce it from memory.

Dunlosky, Rawson, Marsh, Nathan, and Willingham (2013) evaluated ten learning techniques across four dimensions and rated practice testing as "high utility" — beneficial across ages, materials, and tasks, and relatively easy to implement. This is the closest thing the science of learning has to an unqualified recommendation.

3.2 THE METACOGNITIVE PROBLEM

There is a metacognitive twist to the testing effect that matters enormously for practice. Students consistently misjudge which strategies produce the best learning. Rereading feels productive — the material becomes familiar, creating a fluency illusion — while retrieval practice feels effortful and uncertain. Students therefore gravitate toward the least effective strategies and avoid the most effective ones. Kirk-Johnson, Galla, and Fraundorf (2019) investigated this further and found evidence for what they call the "misinterpreted effort" hypothesis: learners interpret the effort of retrieval practice as a signal that the strategy is not working, when in fact the effort is the signal that it *is* working.

This metacognitive error is not a minor curiosity; it is one of the most practically important findings in the field. If students left to their own devices will systematically avoid the most effective learning strategy because it feels difficult, then merely informing students about retrieval practice is insufficient. The curriculum itself must build retrieval practice into its structure, so that students engage in it whether or not they believe it is working.

3.3 FROM THE LABORATORY TO THE CLASSROOM

The translation of retrieval practice from laboratory to classroom has been one of the success stories of cognitive science education research, but the story is more complicated than it first appears.

Agarwal, Nunes, and Blunt (2021) conducted the most comprehensive systematic review of retrieval practice in authentic school settings. Their review confirms that retrieval practice works in classrooms — not just in laboratories with undergraduate psychology students and word lists. The effect generalizes across ages (elementary through college), subjects (science, social studies, languages), and assessment types. This is genuine and important.

McDermott (2020), writing in the *Annual Review of Psychology*, emphasized that retrieval practice also improves transfer — the ability to apply knowledge in new contexts — though the transfer benefits are less consistent and smaller than the retention benefits. This caveat matters. Retrieval practice is an excellent tool for building durable factual and conceptual knowledge, but it is not a panacea for all learning goals.

However, Bego, Chastain, Pyles, and DeCaro (2024) found that effect sizes for spaced retrieval practice in authentic STEM courses were smaller than laboratory studies had predicted. This is not a failure of the technique; rather, it illustrates a general principle that should temper expectations: laboratory effect sizes are typically larger than classroom effect sizes because laboratories control for variables — motivation, prior knowledge, competing demands on attention — that classrooms cannot control. The "glass half full" interpretation is that retrieval practice works in real classrooms. The "glass half empty" interpretation is that the effects may be modest enough that a practitioner could implement retrieval practice faithfully and see only small improvements, especially if other factors (student motivation, course design, assessment alignment) are not also addressed.

3.4 BOUNDARY CONDITIONS: WHEN TESTING DOES NOT HELP

Retrieval practice is not universally beneficial. Several boundary conditions limit its effectiveness:

Material complexity. Retrieval practice is most effective for factual and conceptual knowledge that can be discretely retrieved — terms, definitions, principles, procedures. Its effectiveness for complex, integrative understanding is less clear. When the goal is deep comprehension of interconnected ideas rather than recall of discrete facts, retrieval practice may need to be combined with other strategies (elaboration, self-explanation, worked examples) to be effective.

Feedback requirements. Retrieval practice without feedback can actually reinforce errors if students consistently retrieve incorrect information. The benefit of retrieval practice depends on corrective feedback — either immediately or with a short delay — to ensure that the retrieved information is accurate. In classroom implementations, this means that low-stakes quizzing must include answer review, not just scoring.

Prior knowledge interactions. Students with very low prior knowledge may not have enough to retrieve, making retrieval practice frustrating rather than productive. Retrieval practice presupposes that something has been initially encoded; it is a strengthening mechanism, not an encoding mechanism. For genuine novices encountering entirely new material, initial instruction (worked examples, direct explanation) must precede retrieval practice.

Type of assessment. The benefit of retrieval practice is largest when the criterion test requires the same type of retrieval as the practice. Practicing with flashcards prepares students well for fact-recall tests but less well for application or transfer tasks. This is not surprising, but it is often overlooked in implementations that rely heavily on one format of retrieval practice.

The initial learning requirement. Retrieval practice strengthens existing memories; it does not create them. A common implementation error is to introduce retrieval practice too early — before students have had adequate initial exposure to the material. A quiz on material that has barely been introduced does not produce the testing effect; it produces confusion and frustration. The sequence matters: initial encoding (through instruction, reading, worked examples, or productive failure) must precede retrieval practice. The testing effect operates on memories that exist but are fragile; it cannot operate on memories that have not yet been formed.

These boundary conditions, taken together, paint a more nuanced picture than the simple recommendation to "use retrieval practice." The technique is powerful, but its power depends on implementation details: adequate initial encoding, corrective feedback, varied retrieval formats matched to desired outcomes, and sufficient prior knowledge. A curriculum that builds in well-designed retrieval practice — with these conditions met — can expect meaningful improvements in retention. A curriculum that implements retrieval practice poorly — without feedback, before initial encoding, or using formats mismatched to learning goals — may see little benefit and may even produce harmful reinforcement of errors.

4 SPACING AND INTERLEAVING

4.1 THE SPACING EFFECT

The spacing effect — the finding that learning is more durable when practice is distributed over time rather than massed into a single session — is one of the oldest findings in experimental psychology, dating to Ebbinghaus (1885). It has been confirmed in hundreds of studies across more than a century and is not seriously contested by anyone.

The practical question is not whether spacing works — it does — but how to translate optimal spacing intervals into curriculum schedules. The most cited heuristic comes from the laboratory literature: the optimal gap between practice sessions should be approximately 10–30% of the desired retention interval. If you want to retain information for one month, space your practice sessions three to nine days apart. If you want to retain for one year, space them one to three months apart.

This heuristic is useful but imprecise, and its application to real curricula is fraught with practical challenges. A university course that meets three times per week for fifteen weeks does not have the flexibility to space practice of early material over months. A K-12 curriculum that covers dozens of topics per year cannot easily implement individualized spacing schedules for each student and each topic. The spacing effect tells us that massed practice (cramming) is suboptimal and that distributed practice is better — but the optimal distribution depends on variables (desired retention interval, material complexity, individual differences) that curriculum designers cannot easily control.

The mechanism underlying the spacing effect is also worth understanding, because it explains why spacing feels counterproductive to learners. When practice is spaced, the learner partially forgets the material between sessions. This partial forgetting is not a bug; it is the feature. Re-retrieving material that has been partially forgotten strengthens the memory trace more than re-encountering material that is still fresh. The effort of retrieval — the sense of struggle to recall something that was once known — is the signal that consolidation is occurring. This connects directly to the desirable difficulties framework: spacing creates difficulty through forgetting, and that difficulty is desirable because it drives deeper consolidation.

The spacing effect also interacts with the testing effect in a way that has important practical implications. Spaced retrieval practice — combining spacing with active retrieval — is more effective than either spaced restudying or massed retrieval practice. The combination works because spacing introduces the beneficial forgetting that makes retrieval effortful, and retrieval practice capitalizes on that effort to strengthen memory traces. This is why "cramming" — massed restudying the night before an exam — produces the worst of all worlds: it eliminates the spacing benefit and uses the least effective study strategy (rereading) to do it.

A stubborn practical problem remains: students overwhelmingly prefer massed practice because it produces a fluency illusion. When students reread material shortly after studying it, they recognize it easily and conclude they have learned it well. When students attempt to retrieve material after a delay, they struggle and conclude they have learned it poorly. Both conclusions are wrong — fluency is not learning, and retrieval difficulty is not evidence of failure — but the metacognitive error is powerful and persistent. This is another instance of the misinterpreted-effort problem

identified by Kirk-Johnson et al. (2019) in the context of retrieval practice, and it has the same practical implication: curriculum designers cannot rely on students to space their own practice. The spacing must be built into the curriculum structure.

4.2 THE INTERLEAVING EFFECT

The interleaving effect — the finding that mixing different types of problems or topics during practice produces better learning than studying each type in a separate block — is younger and more nuanced than the spacing effect. Brunmair and Richter (2019) conducted a definitive meta-analysis and found that interleaving reliably improves learning, but with important moderators. The effect is strongest when:

1. The interleaved categories are similar to each other (making discrimination challenging)
2. The criterion test requires discriminating between categories (identifying which approach to use)
3. The material involves inductive category learning (learning to classify rather than learning facts)

The mechanism is also reasonably well understood. Birnbaum, Kornell, Bjork, and Bjork (2012) demonstrated that interleaving enhances inductive learning through two complementary processes: discrimination (contrasting different categories makes their distinguishing features salient) and retrieval (switching between categories forces retrieval of each category's features from memory). Blocked practice allows learners to use a single strategy repeatedly without ever deciding which strategy is appropriate, while interleaving forces learners to practice the act of discrimination — identifying which type of problem they are facing and selecting the appropriate solution approach. This discrimination skill is never exercised in blocked practice, which is why blocked practice feels more fluent and productive but produces worse long-term learning.

4.3 WHY SPACING AND INTERLEAVING ARE NOT THE SAME THING

A common error — both in popular science writing and in some educational implementation guides — is to conflate spacing and interleaving. They are often discussed together, and their practical effects can overlap (interleaving inherently introduces some spacing between instances of the same type). But Chen, Paas, and Sweller (2021) argued that they operate through fundamentally different mechanisms and should be theoretically distinguished.

Spacing works primarily through memory consolidation — distributing practice over time allows forgetting to occur between sessions, and the act of re-retrieving partially forgotten information strengthens the memory trace. The mechanism is about the time interval and the forgetting-retrieval cycle.

Interleaving works primarily through discrimination — mixing categories forces the learner to compare and contrast different types of problems, developing the ability to identify relevant features and select appropriate strategies. The mechanism is about the juxtaposition of different types of material.

This distinction has practical implications. Spacing is beneficial for all types of learning where long-term retention is the goal. Interleaving is particularly beneficial when the learning goal involves discriminating between similar categories or selecting the appropriate approach from

among several options. A curriculum designer who uses spacing to distribute retrieval practice over time and interleaving to mix similar problem types during practice sessions is using both effects, but for different purposes.

4.4 PRACTICAL IMPLEMENTATION CHALLENGES

The practical challenges of implementing spacing and interleaving in real curricula are substantial. Most curricula are organized by topic — a unit on fractions, followed by a unit on decimals, followed by a unit on percentages — which is optimal for blocked practice and terrible for both spacing and interleaving. Implementing spaced, interleaved practice requires either reorganizing the curriculum (difficult, and resisted by teachers and textbook publishers) or supplementing the standard curriculum with additional practice opportunities that deliberately mix and space material (feasible, but requires additional time and resources).

The most practical approach, supported by the evidence, is to build cumulative review into the curriculum — regular practice sessions that draw on all previously covered material, not just the most recently taught topic. This naturally introduces both spacing (earlier material is revisited over time) and interleaving (different types of material are mixed within the same practice session). The challenge is that cumulative review requires careful curation to avoid overwhelming students with too much material, and it requires ongoing assessment to determine which material each student most needs to revisit.

5 THE EXPERTISE REVERSAL EFFECT

5.1 THE CORE FINDING

The expertise reversal effect is arguably the single most important boundary condition in cognitive science of learning, because it demonstrates that there is no universally optimal instructional approach. What works for novices can harm experts, and what works for experts can harm novices.

Kalyuga, Ayres, Chandler, and Sweller (2003) — a paper with a field-weighted citation impact of 41.14, reflecting its influence — documented this effect across multiple CLT-derived instructional techniques. Worked examples help novices but impede experts, who benefit more from problem-solving practice. Detailed guidance helps novices but becomes redundant noise for experts whose schemas already incorporate that information. Integrated formats help novices who need the integration but harm experts for whom the integrated information is already mentally connected.

The mechanism is straightforward within CLT: as learners develop expertise, they build schemas in long-term memory that organize what was once many separate elements into single chunks. Instructional techniques designed for novices work by reducing the cognitive load of processing many separate elements — but for experts who have already chunked those elements, the same techniques impose extraneous load by requiring the expert to process instructional support that is now redundant.

5.2 THE 2025 META-ANALYSIS: WHAT WE NOW KNOW

Tetzlaff, Simonsmeier, Peters, and Brod (2025) conducted the first comprehensive meta-analysis of the expertise reversal effect, synthesizing 176 effect sizes from 60 experimental studies involving 5,924 participants. Their findings are important for both confirming the effect and revealing its nuances.

The effect is robust. Low prior knowledge learners learn better from high-assistance instruction ($d = 0.505$). High prior knowledge learners learn better from low-assistance instruction ($d = -0.428$). These are medium effect sizes, large enough to matter practically.

The effect is asymmetric. Providing assistance to novices has a stronger effect than withholding assistance from experts. This asymmetry has a practical implication that deserves emphasis: if a curriculum designer must choose between providing too much scaffolding or too little, erring on the side of more scaffolding is less harmful than erring on the side of less. The cost of giving experts unnecessary guidance (some redundancy processing, some boredom) is lower than the cost of giving novices insufficient guidance (cognitive overload, confusion, disengagement).

The effect is moderated. Three moderators emerged as significant:

Type of prior knowledge assessment: The effect is stronger when prior knowledge is measured with performance-based assessments than with self-report measures. This is not surprising — performance measures are more accurate than self-reports — but it matters for practice because effective adaptive instruction depends on accurate diagnosis of learner knowledge.

Educational status: The effect is more robust for university students and older learners than for younger students. For younger students, the evidence is less clear. This may reflect the fact that

younger learners generally have lower prior knowledge across all topics, making it harder to find genuine "experts" for comparison.

Content domain: The effect is well-established in STEM domains but less clear in humanities and language learning. This is a significant boundary condition. Humanities and language learning involve different types of knowledge structures (more contextual, more interpretive, less hierarchically organized) that may not interact with instructional scaffolding in the same way as STEM content.

5.3 WHAT THE EXPERTISE REVERSAL EFFECT MEANS FOR PRACTICE

The practical implication of the expertise reversal effect is that effective instruction must adapt to the learner's current knowledge state. This is not a subtle recommendation; it is a fundamental requirement. A curriculum that delivers the same instruction to all learners — regardless of their prior knowledge — is guaranteed to be suboptimal for most of them. It will provide too much scaffolding for advanced learners and too little for novices.

Tetzlaff, Schmiedek, and Brod (2020) proposed a dynamic framework for developing personalized education that takes the expertise reversal effect as a cornerstone. Their framework argues that adaptive instruction requires continuous assessment of learner knowledge, dynamic adjustment of instructional support, and gradual fading of scaffolding as expertise develops. This is conceptually straightforward but operationally demanding — it requires assessment systems that can diagnose knowledge state in real time and instructional materials that can be dynamically adjusted.

The expertise reversal effect also has implications for the debate between direct instruction and discovery learning. Both sides of that debate have tended to argue for the universal superiority of their preferred approach. The expertise reversal effect renders both positions untenable. Direct instruction is superior for novices. Problem-solving and exploratory approaches become superior as expertise develops. The question is not "which is better?" but "which is better for this learner, with this level of prior knowledge, learning this particular material?" This is the question that productive failure research has been addressing, and it is to that topic we now turn.

Part III

FRONTIERS

6 DESIRABLE DIFFICULTIES AND PRODUCTIVE FAILURE

6.1 THE DESIRABLE DIFFICULTIES FRAMEWORK

Robert Bjork's concept of desirable difficulties — the idea that conditions making learning feel harder can improve long-term retention and transfer — creates a productive tension with cognitive load theory. CLT predicts that difficulty impairs learning because it overloads working memory. Bjork's framework predicts that certain types of difficulty enhance learning because they force deeper processing.

The resolution is not that one framework is right and the other wrong, but that they apply to different types of difficulty. Difficulty stemming from poor instructional design (split attention, redundant information, incoherent presentation) is undesirable — it imposes extraneous cognitive load without promoting learning. Difficulty stemming from generative processing (retrieval practice, spacing, interleaving, generation) is desirable — it imposes productive cognitive demands that strengthen memory traces and promote deeper encoding.

But this distinction, clean as it sounds, is difficult to apply in practice. Lodge, Kennedy, Lockyer, Arguel, and Pachman (2018) argued that the metacognitive experience of difficulty feels the same regardless of whether it is desirable or undesirable — the learner experiences confusion, effort, and uncertainty in both cases. This means that learners cannot easily distinguish between productive struggle and unproductive floundering, and neither can teachers who are observing them. The "S2D2" framework (Start and Stick to Desirable Difficulties) proposed by de Bruin and colleagues (2020) attempts to help learners identify and persist through desirable difficulties despite the discomfort, but it is a metacognitive intervention that requires substantial training and may not transfer across contexts.

Kirk-Johnson, Galla, and Fraundorf (2019) provided further evidence for the problem with their "misinterpreted effort" hypothesis. Learners interpret the effort associated with desirable difficulties as a signal that the learning strategy is not working, leading them to abandon effective strategies in favor of less effortful but less effective ones. This is a metacognitive trap: the very features that make a difficulty desirable (effort, challenge, initial failure) are the features that lead learners to reject it.

6.2 PRODUCTIVE FAILURE: THE STRONGEST CASE FOR STRUCTURED DIFFICULTY

Kapur's productive failure research program represents the most sustained and rigorous attempt to harness desirable difficulties in instructional design. The approach inverts the conventional instructional sequence: instead of instruction followed by practice (I → PS), productive failure has learners attempt to solve problems first and receive instruction afterward (PS → I).

Kapur (2025) presents a comprehensive synthesis of this research in book form, drawing on over fifty studies and 160+ comparisons. The core finding is striking: productive failure consistently produces effects on conceptual understanding and transfer that are substantially larger — up to three times larger — than those produced by well-implemented direct instruction, while matching direct instruction on procedural skills. The effect is not driven by poor implementations of direct

instruction in the comparison groups; Kapur is careful to compare productive failure against well-designed, well-implemented direct instruction.

The mechanism, as Kapur articulates it, operates through four processes (the 4A framework):

Activation. Attempting to solve a problem before instruction activates relevant prior knowledge — both correct and incorrect — creating cognitive hooks onto which subsequent instruction can attach. This activation is broader and deeper than what occurs when learners simply listen to a lecture and then practice.

Awareness. The experience of failing to solve the problem creates awareness of knowledge gaps. Learners discover what they do not know, which makes them more receptive to instruction that addresses those gaps. This connects to VanLehn's finding that "impasses" during problem-solving are strongly associated with learning gains.

Affect. The struggle of attempting an unsolved problem creates situational interest and curiosity — what Kapur calls the "learning cliffhanger" effect. The need for closure drives engagement with subsequent instruction. Negative emotions like confusion and frustration, when experienced in a safe context, can be productive rather than debilitating.

Assembly. The instruction phase following the problem-solving attempt serves to assemble the learner's partial solutions and prior knowledge into a coherent understanding. The teacher's role is not merely to present the correct solution but to explicitly compare and contrast the learner's attempts with the canonical solution, helping the learner see where their ideas were partially correct and where they went wrong.

6.3 KAPUR'S CENTRAL INSIGHT: THE BASIC KNOWLEDGE FALLACY

One of Kapur's most important arguments is what he calls the "basic knowledge fallacy" — the assumption that if you teach foundational knowledge efficiently (through clear, direct instruction), you can then build higher-order understanding on top of it. The fallacy is that *how* foundational knowledge is learned affects whether higher-order understanding can be built on it at all. Two learners who have mastered the same procedural skills through different methods may have very different capacities for conceptual understanding and transfer, because one has learned the procedures in a way that also built conceptual structure, while the other has learned them in a way that did not.

This argument, if correct, has profound implications for curriculum design. It means that "covering the basics first" — the standard approach in most curricula — may actually undermine the goal of deeper understanding, not because the basics are unimportant, but because the method of teaching them matters as much as the content itself. Productive failure teaches the same content but through a process that simultaneously builds the conceptual connections necessary for transfer.

6.4 BOUNDARY CONDITIONS OF PRODUCTIVE FAILURE

Productive failure is not universally beneficial. Several boundary conditions have been identified:

Prior knowledge dependence. He, Fiorella, and Lemons (2025) found that the relative effectiveness of problem-solving-first versus instruction-first depends on learners' prior knowledge. For learners with very low prior knowledge, the initial problem-solving phase may be too challenging to be productive — there is not enough relevant prior knowledge to activate, and the struggle becomes frustrating rather than illuminating. This connects productive failure directly to the expertise reversal effect: productive failure may be less effective for genuine novices (who benefit more

from direct instruction) and more effective for learners who have some relevant prior knowledge but have not yet developed full understanding.

This is a critical finding because it complicates the simple narrative that productive failure is always superior to direct instruction. The optimal instructional sequence may depend on the learner's knowledge state — direct instruction for genuine novices, productive failure for learners with intermediate knowledge, and open-ended problem-solving for experts. If confirmed by further research, this would represent a significant refinement of Kapur's claims.

Design quality dependence. Productive failure's effectiveness depends heavily on the quality of the task design and the subsequent instruction phase. Kapur's seven design features (challenging but accessible, contextualized, admitting multiple solutions, affectively engaging, using contrasting cases, varying the cases, minimizing computational load) are necessary conditions, not optional enhancements. A poorly designed productive failure task — one that is too difficult, too abstract, or that does not admit multiple approaches — will produce unproductive failure: frustration without learning.

The instruction phase is equally critical. Simply having students attempt a problem and then giving them the correct answer is not productive failure; it is guessing followed by telling. The instruction must explicitly compare and contrast the students' attempts with the canonical solution, highlighting which features of their approaches were on the right track and which were not. This requires teachers to be responsive to students' actual solutions, not just to deliver a pre-planned lecture.

Domain limitations. Productive failure has been most extensively studied in mathematics and science, where problems have clear structures and canonical solutions that can be compared against student attempts. Its application to humanities, creative work, and other ill-structured domains is less well understood. Kapur himself acknowledges this, noting that productive failure applies to tasks where "concepts can be meaningfully explored" — a qualification that may exclude some types of learning.

The collaboration question. Kapur's meta-analytic work suggests that productive failure is more effective when students collaborate during the problem-solving phase, but this introduces the collaboration costs identified by Kirschner et al. (2018). The optimal implementation may involve individual generation followed by collaborative sharing, which captures the benefits of diverse perspectives without the full transaction costs of sustained collaboration.

6.5 THE RELATIONSHIP TO DESIRABLE DIFFICULTIES

Productive failure is often discussed as an instantiation of Bjork's desirable difficulties framework, and there are clear connections — both involve making learning harder in the short term to improve long-term outcomes. But there are also important differences.

Desirable difficulties (retrieval practice, spacing, interleaving) are primarily memory-enhancement strategies — they improve retention and accessibility of existing knowledge. Productive failure is primarily a comprehension and transfer strategy — it improves conceptual understanding and the ability to apply knowledge in new contexts. The mechanisms are different: desirable difficulties work through memory consolidation and retrieval strengthening; productive failure works through prior knowledge activation, gap awareness, and schema construction.

Fiorella (2023) attempted to bring these perspectives together under the umbrella of "generative learning" — the idea that learning is enhanced when learners actively generate or construct knowledge rather than passively receiving it. This is a useful framework because it encompasses both desirable difficulties (where generation takes the form of retrieval) and productive failure

(where generation takes the form of solution attempts), while also including other generative activities like self-explanation, summarization, and teaching.

7 TRANSFER: THE GREAT UNSOLVED PROBLEM

7.1 WHY TRANSFER MATTERS MORE THAN ANYTHING ELSE

Transfer — the ability to apply knowledge learned in one context to a different context — is the implicit goal of all education. A student who can solve the practice problems at the end of a textbook chapter but cannot apply the same principles in a new context has not learned anything useful. Yet transfer is also the most elusive learning outcome, the one that cognitive science has struggled most to explain and predict.

The Lo survey identified transfer as Gap 1 — the most significant gap in the field's understanding. This investigation confirms that assessment. The cognitive science of transfer is thin relative to its importance, and the practical guidance it offers is limited.

7.2 WHAT WE KNOW

Barnett and Ceci (2002) proposed a useful taxonomy of transfer along multiple dimensions: content (what is transferred — learned skill, principle, or representation), context (physical, social, temporal, functional, and modality differences between learning and transfer), and performance change (speed, accuracy, or approach). Their key insight is that "transfer" is not a single phenomenon but a family of phenomena, and what counts as evidence for transfer depends on how far apart the learning and transfer contexts are along each dimension.

Near transfer — applying knowledge to problems that are structurally similar to the practice problems but differ in surface features — is reliably achieved by cognitive science techniques, particularly when those techniques include interleaving (which develops discrimination skills) and varied practice (which broadens the range of examples encoded in memory). Far transfer — applying knowledge to problems that are structurally different from the practice problems — is far more difficult to achieve and far more sparsely documented.

Kapur's (2025) productive failure research provides some of the strongest evidence for instructional approaches that promote transfer. In his meta-analytic work, productive failure reliably outperforms direct instruction on transfer tasks, with effect sizes up to three times those of direct instruction. The mechanism appears to involve the construction of more flexible knowledge representations — learners who have explored multiple solution approaches during the problem-solving phase develop schemas that are less tied to specific problem formats and more adaptable to new contexts. Kapur's analysis of why this occurs is worth unpacking. When learners attempt to solve a problem before instruction, they generate multiple representations and solution paths. These varied representations encode the problem's deep structure from multiple angles, creating what Kapur calls "encoding variability" — the same concept encoded through different approaches, creating multiple retrieval paths. When instruction subsequently connects these varied representations to the canonical solution, the learner's mental model is richer and more flexible than one built through a single instructional pathway. This flexibility is precisely what transfer requires: a representation that is not locked to the specific format in which it was learned.

The implication is important: transfer may not be primarily about abstracting away from specifics (the traditional view) but about accumulating diverse, specific encounters that, collectively, reveal

the underlying structure. This is consistent with the interleaving research, where encountering the same category in varied contexts builds discriminative ability — a form of transfer within the category discrimination task.

7.3 WHAT WE DO NOT KNOW

The honest assessment is that the field does not have a reliable, well-tested theory of far transfer. We know that some instructional approaches produce more transfer than others. We know that interleaving helps with discrimination transfer, that productive failure helps with conceptual transfer, and that retrieval practice helps with retention-based transfer. But we do not have a general theory that predicts when transfer will occur, under what conditions, and to what extent.

The history of transfer research is, frankly, discouraging. For over a century, researchers have hoped to demonstrate that education produces general cognitive improvement — that studying Latin improves logical reasoning, that learning programming improves problem-solving, that training in one domain makes learners broadly smarter. The evidence for such general transfer is weak at best. Thorndike and Woodworth's "identical elements" theory, proposed in 1901, suggested that transfer occurs only to the extent that the original and transfer tasks share common elements — and subsequent research has largely confirmed this conservative position. Transfer is specific, not general. It follows the contours of shared structure between tasks, and the amount of transfer decreases rapidly as the structural similarity between learning and transfer contexts decreases.

Several specific gaps are worth noting:

The role of abstract representation. A common hypothesis is that transfer depends on the learner constructing abstract representations that capture the deep structure of a problem independent of its surface features. This hypothesis is plausible and has some support, but the process by which learners construct such representations — and how instruction can facilitate that construction — is not well understood.

Individual differences. Some learners transfer more readily than others, and the sources of this variation are poorly understood. Prior knowledge, working memory capacity, metacognitive skill, and motivation have all been proposed as factors, but the relative importance of each and their interactions are not well characterized. De Lima and Buratto (2024) began investigating individual differences in retrieval practice, but the broader question of who benefits most from transfer-promoting instruction remains open.

Time scale. Most transfer studies measure performance days or weeks after learning. Whether the transfer-promoting effects of interleaving, productive failure, and other techniques persist over months or years is largely unknown. Curriculum designers need to know not just whether a technique promotes transfer immediately, but whether the transfer benefits are durable over the time scales relevant to education.

Domain specificity. Transfer appears to be highly domain-specific — expertise in chess does not transfer to general reasoning, and expertise in one branch of mathematics does not automatically transfer to others. This suggests that the cognitive structures built during learning are more contextualized than we might hope. If transfer is inherently limited by domain boundaries, then the educational goal of producing flexible, broadly applicable thinkers is more challenging than commonly assumed.

8 COGNITIVE SCIENCE IN ILL-STRUCTURED DOMAINS

8.1 THE GAP THE FIELD HAS NOT FILLED

Most cognitive load theory research — and most retrieval practice research, most spacing research, and most productive failure research — uses well-structured problems in mathematics and science. These are problems with clear initial states, clear goal states, and clear operators for moving from one to the other. The learner may not know the answer, but the answer exists and is unambiguous.

Much of the most important learning in education occurs in ill-structured domains, where there is no single correct answer and the criteria for evaluating solutions are themselves contested. Writing a persuasive essay, analyzing a historical event, designing a product, navigating an ethical dilemma, developing a business strategy — these are all ill-structured tasks where CLT's prescriptions become unclear.

The Lo survey flagged this as Gap 5, and this investigation confirms that the gap has not been significantly narrowed since 2019. The literature on cognitive load in ill-structured domains remains thin.

8.2 WHAT LITTLE WE KNOW

The 4C/ID model. Van Merriënboer and colleagues' four-component instructional design model (4C/ID) is the most serious attempt to extend CLT-derived principles to complex, real-world tasks. The model addresses whole-task learning — tasks that cannot be meaningfully decomposed into independent sub-skills without losing their essential character. It proposes four components: learning tasks (whole, authentic tasks of decreasing scaffolding), supportive information (mental models and cognitive strategies), procedural information (just-in-time instruction for routine aspects), and part-task practice (for automating critical sub-skills).

The 4C/ID model is conceptually sophisticated and has been applied in fields like medicine, engineering, and teacher training. But it is primarily a design framework rather than an empirically tested theory. Its specific predictions about cognitive load in complex tasks have not been tested as rigorously as CLT's predictions about well-structured problems.

Self-regulation and cognitive load. Nückles, Roelle, Glogger-Frey, Waldeyer, and Renkl (2020) investigated cognitive load in writing-to-learn tasks — a paradigm that combines elements of both well-structured (following genre conventions) and ill-structured (generating original arguments) processing. They found that self-regulation plays a much larger role in managing cognitive load during writing than during well-structured problem-solving, because the learner must not only process information but also make strategic decisions about what to write, how to organize it, and when to revise. This suggests that CLT's standard prescriptions (reduce extraneous load, manage intrinsic load) need to be supplemented with self-regulatory scaffolding when tasks become ill-structured.

Collaborative approaches. Kirschner, Sweller, Kirschner, and Zambrano's (2018) extension of CLT to collaborative learning is relevant here because ill-structured problems often require collaboration — they are too complex for individual working memory and benefit from multiple perspectives. The collaborative CLT framework's prediction that collaboration is beneficial

when task complexity exceeds individual capacity may be particularly applicable to ill-structured domains, where the complexity is inherently high.

8.3 WHY THE GAP MATTERS

The practical importance of this gap cannot be overstated. A curriculum designer working in STEM has access to a rich evidence base about how to sequence problems, when to use worked examples versus practice, how to space and interleave, and when to fade scaffolding. A curriculum designer working in humanities, social sciences, design, or creative writing has almost none of this guidance.

This does not mean that cognitive science principles are irrelevant in ill-structured domains — working memory is still limited, prior knowledge still matters, and retrieval still strengthens memory. But the specific instructional prescriptions derived from CLT have not been tested in these contexts, and it would be irresponsible to assume that findings from mathematics and science transfer directly to essay writing or ethical reasoning.

There is a deeper issue here that goes beyond simply testing existing CLT effects in new domains. Ill-structured problems may differ from well-structured problems not merely in degree but in kind. In a well-structured problem, the elements to be processed are definable, and the relationships between them are governed by rules (mathematical operations, physical laws, grammatical structures). The learner's task is to build a schema that captures these rules. In an ill-structured problem, the elements themselves may be ambiguous (what counts as a relevant historical fact?), the relationships between them may be contestable (was the economic downturn a cause or a consequence of the political upheaval?), and the criteria for a good solution may depend on values, audience, and purpose (is this a good essay?). It is not clear that the concept of element interactivity — counting the number of elements that must be processed simultaneously — can be meaningfully applied to tasks where what counts as an "element" is itself a judgment call.

This may explain why the expertise reversal effect is weaker in humanities and language learning (Tetzlaff et al., 2025). In well-structured domains, expertise means having schemas that chunk multiple elements into single units. In ill-structured domains, expertise may mean something different — perhaps the ability to hold multiple interpretations in productive tension, to recognize relevant contextual factors, or to apply flexible judgment. If expertise operates differently, then the CLT mechanism that generates the expertise reversal effect (schema-based chunking that makes instructional support redundant) may not apply in the same way.

The questions that need answering include: What constitutes a "worked example" in essay writing or historical analysis? What counts as "element interactivity" when the elements are interpretive claims rather than mathematical operations? How should scaffolding be faded in domains where there is no single correct solution toward which the learner is converging? How does the expertise reversal effect manifest when expertise itself is defined differently (as interpretive sophistication rather than as problem-solving efficiency)?

One promising direction is productive failure, which may be more naturally suited to ill-structured domains than standard CLT prescriptions. Productive failure's emphasis on generating multiple solutions, comparing and contrasting approaches, and building understanding through exploration aligns well with the kind of thinking that ill-structured domains require. But this is speculative — productive failure has been primarily studied in mathematics, and its application to humanities and creative work remains untested.

These are not merely academic questions. They determine whether cognitive science can guide curriculum design across the full range of educational domains, or only within the well-structured STEM domains where it has been primarily studied.

Part IV

SYNTHESIS

9 INTEGRATION: HOW THE PIECES FIT TOGETHER

9.1 REINFORCEMENTS AND TENSIONS

The sub-topics reviewed above are not independent. They interact in ways that are sometimes reinforcing and sometimes in tension, and understanding these interactions is essential for coherent curriculum design.

Retrieval practice and spacing reinforce each other. Spaced retrieval practice — combining the testing effect with distributed practice — produces stronger learning than either technique alone. This is one of the field's most actionable findings: building regular, spaced quizzing into a curriculum is an evidence-based practice that leverages two robust effects simultaneously.

Interleaving and retrieval practice reinforce each other. Sana and Yan (2022) found that interleaving retrieval practice promotes science learning more effectively than either technique alone. Mixing different types of practice problems during retrieval sessions forces learners to discriminate between problem types (the interleaving benefit) while also strengthening memory traces through retrieval (the testing benefit).

Productive failure and CLT create a tension — but a productive one. CLT predicts that novice learners should receive maximum instructional support to reduce cognitive load. Productive failure deliberately withholds support and allows learners to struggle. The resolution, as discussed above, is that these apply to different types of difficulty and potentially to different levels of prior knowledge. But the tension is real, and it has practical implications: a curriculum designer who follows CLT prescriptions will design differently than one who follows productive failure prescriptions, and neither is wrong in all cases.

The productive failure–CLT tension can be partially resolved through Kapur's (2025) concept of the "basic knowledge fallacy." CLT assumes that reducing cognitive load during initial learning is always beneficial. Kapur argues that the *way* foundational knowledge is learned — not just whether it is learned — affects what can be built on it later. A learner who acquires a procedure through worked examples may have a schema that supports procedural fluency but not conceptual flexibility. A learner who acquires the same procedure through productive failure may have a schema that supports both. If this is correct, then CLT's prescriptions optimize for initial learning efficiency at the potential cost of later transfer and deep understanding. The practical resolution is not to choose between CLT and productive failure but to understand which goals each serves: CLT for efficient procedural skill acquisition, productive failure for deep conceptual understanding and transfer.

The expertise reversal effect connects everything. The expertise reversal effect is not merely a boundary condition on the worked example effect; it is a general principle that applies to all the techniques discussed in this review. Retrieval practice may be less beneficial for complete novices than for learners who have some initial knowledge to retrieve. Interleaving may be less beneficial when learners have not yet developed the baseline knowledge needed to discriminate between categories. Productive failure may be less beneficial for genuine novices who lack the prior knowledge to generate meaningful (even if incorrect) solutions. The expertise reversal effect means that the optimal instructional strategy shifts as the learner develops, and that static curricula

— curricula that deliver the same instruction to all learners regardless of knowledge state — are systematically suboptimal.

9.2 THE MISSING INTEGRATION: MOTIVATION AND COGNITION

One of the most significant gaps in the cognitive science of learning is the near-total absence of integration with motivational science. Cognitive science treats learning as a computational process: encoding, storage, retrieval, schema construction. Motivational science treats learning as a goal-directed, emotionally laden activity: autonomy, competence, relatedness, interest, self-efficacy.

Both perspectives have strong evidence. Neither alone accounts for how people actually learn. A student who understands retrieval practice is effective but lacks the motivation to practice will not benefit from the technique. A student who is highly motivated but uses ineffective strategies (highlighting, rereading) will learn less than they should. The intersection of motivation and cognitive strategy — how to get students to use effective strategies and to persist through the discomfort they produce — is underresearched and deserves focused attention.

Kapur's productive failure work begins to address this intersection, because the 4A framework explicitly includes affect as a mechanism. The experience of struggle, curiosity, and the "learning cliffhanger" are motivational states that productive failure deliberately cultivates. But this integration is still incomplete — productive failure does not address what happens when students' motivational orientation (performance-avoidance, for instance) makes them unwilling to engage in the initial problem-solving phase at all.

10

THE TRANSLATION PROBLEM: FROM LABORATORY TO CLASSROOM

10.1 WHAT IS LOST

Every finding reviewed in this dissertation was initially demonstrated in a laboratory. Some have been successfully translated to classroom settings. Others have not been tested in classrooms at all. And those that have been tested in classrooms consistently show smaller effect sizes than laboratory demonstrations.

This is not a bug in the research; it is a fundamental feature of the difference between laboratories and classrooms. Laboratories control for variables that classrooms cannot: motivation (participants are typically compensated and compliant), prior knowledge (participants can be screened for homogeneity), competing demands (participants focus exclusively on the experimental task), and implementation fidelity (the intervention is delivered exactly as designed).

Classrooms introduce all of these variables simultaneously. Students differ in motivation, prior knowledge, attention, emotional state, peer relationships, and a hundred other factors that interact with instructional techniques in complex ways. A technique that produces a large effect in a laboratory, where confounding variables are controlled, may produce a small effect in a classroom, where confounding variables are abundant — not because the technique is less effective, but because its effect is diluted by uncontrolled variation.

10.2 WHAT COGNITIVE SCIENCE TYPICALLY IGNORES

Several factors that are central to classroom learning are largely absent from cognitive science research:

Motivational variation. Cognitive science assumes that learners are willing to engage with the instructional materials. In classrooms, this assumption frequently fails. A student who does not see the point of the material, who is anxious about failure, or who is distracted by social dynamics will not benefit from optimally designed retrieval practice any more than a car with an empty gas tank will benefit from optimal tire pressure.

Teacher implementation. The effectiveness of any instructional technique depends on how well it is implemented. Productive failure, for instance, requires teachers to respond to students' actual solution attempts during the instruction phase — a demanding skill that not all teachers possess. The gap between "what should be done" (as described in research) and "what is actually done" (as implemented in classrooms) is a major source of the science-practice gap.

Classroom dynamics. Learning does not happen in a social vacuum. A student's willingness to engage in productive failure depends on whether the classroom culture supports risk-taking. A student's response to retrieval practice depends on whether low scores are treated as learning opportunities or as marks of inadequacy. These social-contextual factors are rarely studied in cognitive science but powerfully influence learning outcomes.

Cumulative effects. Most cognitive science studies measure the effect of a single technique over a short period — a few weeks at most. Curriculum designers need to know about cumulative effects over months and years: does spaced retrieval practice produce accelerating benefits as knowledge

accumulates, or do the effects plateau? Does productive failure become more effective as students develop the metacognitive skills to navigate it, or does its effectiveness diminish as the novelty wears off? These questions are largely unanswered.

10.3 THE TRANSLATION IS WORTH DOING

Despite these challenges, the translation of cognitive science to classroom practice is worth pursuing, for two reasons.

First, the core findings are robust enough that even with diluted effect sizes, they produce meaningful improvements. Retrieval practice with an effect size of $d = 0.3$ in a classroom is worth implementing, even if the laboratory effect size was $d = 0.7$. Spacing with modest effect sizes is still superior to cramming with no benefit.

Second, the cumulative effect of implementing multiple evidence-based techniques simultaneously may be larger than any single technique alone. Koedinger, Corbett, and Perfetti (2012) proposed the Knowledge-Learning-Instruction framework as one attempt to bridge this gap systematically, linking different types of knowledge to appropriate learning events and instructional strategies. A curriculum that builds in spaced retrieval practice, uses worked examples that fade into problem-solving as expertise develops, interleaves related topics, and incorporates productive failure for conceptual material is leveraging multiple robust effects. The interactions between these techniques are not well studied, but there is no reason to expect them to cancel each other out, and some reason to expect synergistic effects.

11 CLOSING ASSESSMENT: WHAT A CURRICULUM DESIGNER CAN RELY ON

11.1 THE CONFIDENT RECOMMENDATIONS

The following recommendations are supported by strong, replicated evidence and can be implemented with reasonable confidence:

Build retrieval practice into the curriculum structure. Do not rely on students to test themselves; they will not. Build low-stakes quizzing, recall exercises, and retrieval opportunities into every unit. Include corrective feedback. Vary the format of retrieval to match the desired outcome (fact recall for factual knowledge, application problems for procedural knowledge).

Space practice over time. Distribute practice across days and weeks rather than massing it into single sessions. Build cumulative review into the schedule. When returning to previously covered material, use retrieval practice rather than re-exposition.

Use worked examples for novices, and fade them as expertise develops. For new material, provide fully worked examples that model the solution process. As learners develop competence, gradually remove steps from the examples (faded examples), transitioning to independent problem-solving. Monitor for the expertise reversal — do not continue providing worked examples to learners who no longer need them.

Interleave similar problem types during practice. When students are practicing skills that require discriminating between similar approaches (which formula to apply, which method to use), mix the problem types rather than blocking them. This develops the discriminative skill that blocked practice never exercises.

Integrate related information physically. Place labels on diagrams. Synchronize narration with animation. Do not require learners to mentally integrate information that could be physically integrated in the instructional materials.

Adapt instruction to learner knowledge state. The expertise reversal effect means that one-size-fits-all instruction is guaranteed to be wrong for most learners. Develop mechanisms for assessing prior knowledge and adjusting scaffolding accordingly. This is operationally difficult but cognitively necessary.

11.2 THE CAUTIOUS RECOMMENDATIONS

The following recommendations are supported by evidence but have important boundary conditions that practitioners need to understand:

Use productive failure for conceptual understanding — but not for everything. Productive failure is well-suited for material where conceptual understanding and transfer are the primary goals, where the task admits multiple solution approaches, and where learners have enough prior knowledge to generate meaningful (even if incorrect) attempts. It is less well-suited for genuinely novel material where learners lack any relevant prior knowledge, for purely procedural skills, or for contexts where the classroom culture does not support risk-taking.

Combine cognitive strategies with motivational support. Effective learning strategies like retrieval practice and spacing are effortful and often feel unpleasant. Students who are not supported

in understanding why these strategies work and in persisting through the discomfort may abandon them. Metacognitive instruction — teaching students how to learn and why effective strategies feel harder — is a necessary complement to cognitive strategy implementation.

Be cautious about extrapolating from STEM to other domains. Most of the evidence reviewed in this dissertation comes from mathematics, science, and technical domains. The principles may apply in humanities and creative fields, but they have not been adequately tested there. The expertise reversal meta-analysis (Tetzlaff et al., 2025) found weaker effects in humanities and language learning, suggesting that direct extrapolation may not be warranted.

11.3 WHAT WE CANNOT YET RECOMMEND

Optimal spacing schedules for specific content. The 10–30% heuristic is a rough guide, but cognitive science cannot yet prescribe specific spacing intervals for specific types of content in specific curriculum contexts. Curriculum designers should space practice but should not expect research to specify the ideal spacing with precision.

How to manage cognitive load in ill-structured domains. The field does not have evidence-based guidance for instructional design in domains where problems are open-ended, criteria are contested, and there is no single correct answer. Working memory limitations still apply, but the specific CLT prescriptions (use worked examples, reduce split attention, etc.) have not been validated outside of well-structured domains.

How to promote far transfer reliably. Near transfer can be promoted through interleaving, varied practice, and productive failure. Far transfer — applying knowledge across genuinely different domains — remains elusive. Curriculum designers should not promise far transfer from any instructional approach, because the evidence does not support such promises.

How individual differences interact with instructional techniques. Beyond the expertise reversal effect (which addresses differences in prior knowledge), we know very little about how other individual differences — working memory capacity, motivation, metacognitive skill, personality — interact with specific instructional techniques. The same technique may work differently for different learners in ways that go beyond their knowledge level, and the field has barely begun to investigate these interactions.

11.4 THE QUESTIONS THAT REMAIN

This review has attempted to map the cognitive science of learning as it actually stands — with its strengths, its boundary conditions, and its gaps. The field's strengths are genuine: several well-replicated findings provide actionable guidance for curriculum design. But the gaps are also genuine, and they cluster around the questions that matter most for education as a whole.

The deepest unanswered question is whether the cognitive science framework — with its focus on individual cognition, memory mechanisms, and well-structured problems — can ever fully capture how people learn in the messy, social, motivated, emotionally laden contexts where real learning occurs. The framework is not wrong; it is incomplete. Completing it will require integration with motivational science, sociocultural theory, and the study of learning in domains where there is no clear right answer.

There is a meta-lesson in this review that bears stating explicitly. The cognitive science of learning is at its strongest when it operates at the level of basic mechanisms — retrieval strengthens memory, working memory is limited, prior knowledge moderates instructional effectiveness. These findings are robust because they describe the architecture of human cognition, not the specifics of any

particular educational context. They hold regardless of subject matter, institutional setting, or cultural context, because they reflect how brains process information.

The field is at its weakest when it tries to prescribe specific instructional sequences — "use worked examples for exactly three problems, then switch to faded examples, then to problem-solving" — because these prescriptions depend on context-specific variables (learner knowledge state, material complexity, classroom dynamics) that the research cannot specify for every situation. The gap between general mechanism and specific prescription is the gap that practitioners must bridge, and the cognitive science literature provides the raw materials for that bridge without constructing it.

The most productive stance for a curriculum designer, therefore, is not to treat cognitive science findings as recipes but as constraints and affordances. Working memory is limited — this is a constraint. Retrieval strengthens memory — this is an affordance. The expertise reversal effect means instruction must adapt — this is both a constraint (one-size-fits-all will not work) and an affordance (the direction of adaptation is specified). Productive failure promotes transfer — this is an affordance. But productive failure requires specific design features and sufficient prior knowledge — these are constraints on the affordance. The art of curriculum design lies in optimizing within these constraints while exploiting these affordances, in the specific context of specific learners learning specific material for specific purposes.

For Applied Pedagogy, the practical message is clear: use the cognitive science evidence. It is the strongest evidence available about how to design instruction. But do not treat it as a complete guide. Supplement it with attention to motivation, culture, and domain-specific knowledge. Be honest about the boundaries. And invest in testing these principles in the specific contexts where Applied Pedagogy operates, because the laboratory-to-classroom translation must ultimately be done by practitioners, not by researchers.

BIBLIOGRAPHY

Agarwal, P. K., Nunes, L. D., & Blunt, J. R. (2021). Retrieval practice consistently benefits student learning: A systematic review of applied research in school classrooms. *Educational Psychology Review*, 33(4), 1409–1453.

Ayres, P., Lee, H., Paas, F., & van Merriënboer, J. J. G. (2021). The validity of physiological measures to identify differences in intrinsic cognitive load. *Frontiers in Psychology*, 12, 702538.

Barbieri, C., Miller-Cotto, D., Clerjuste, S. N., & Chawla, K. (2023). A meta-analysis of the worked examples effect on mathematics performance. *Educational Psychology Review*, 35, 51.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, 128(4), 612–637.

Baumgartner, T., Rieser, S., & Stark, R. (2025). Problem-solving before instruction in university mathematics: A replication and extension. *Instructional Science*, 53, 119–143.

Bego, C. R., Chastain, R. J., Pyles, L. M., & DeCaro, M. S. (2024). Spaced retrieval practice in a STEM course: Glass half full or half empty? *Learning and Instruction*, 89, 101837.

Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2012). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & Cognition*, 41, 392–402.

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*, 145(11), 1029–1052.

Castro-Alonso, J. C., de Koning, B. B., Fiorella, L., & Paas, F. (2021). Five strategies for optimizing instructional materials: Instructor- and learner-managed cognitive load. *Educational Psychology Review*, 33(4), 1379–1407.

Chen, O., Paas, F., & Sweller, J. (2021). Spacing and interleaving effects require distinct theoretical bases: A cognitive load theory perspective. *Educational Psychology Review*, 33(4), 1535–1557.

Chen, O., Paas, F., & Sweller, J. (2023). A cognitive load theory approach to defining and measuring task complexity through element interactivity. *Educational Psychology Review*, 35, 63.

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 Bruin, A., Roelle, J., Carpenter, S., & Baars, M. (2020). Synthesizing cognitive load and self-regulation theory: A theoretical framework and research agenda. *Educational Psychology Review*, 32(4), 903–915.

de Jong, T. (2009). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science*, 38(2), 105–134.

de Lima, R. H., & Buratto, L. G. (2024). Individual differences in retrieval practice. *Memory*, 32(5), 567–581.

DeCaro, M. S., DeCaro, D. A., & Rittle-Johnson, B. (2023). Productive failure in online physics problem solving. *Instructional Science*, 51(5), 741–764.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, 14(1), 4–58.

Ebbinghaus, H. (1885). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie* [On Memory: Investigations in Experimental Psychology]. Leipzig: Duncker & Humblot.

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

He, L., Fiorella, L., & Lemons, P. P. (2025). Does instruction-first or problem-solving-first lead to better learning outcomes? It depends on prior knowledge. *Learning and Instruction*, 96, 102074.

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

Kapur, M. (2025). *Productive Failure: Unlocking Deeper Learning Through the Science of Failing*. Jossey-Bass/Wiley.

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

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, 319(5865), 966–968.

Kirk-Johnson, A., Galla, B. M., & Fraundorf, S. H. (2019). Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice. *Cognitive Psychology*, 115, 101237.

Kirschner, P. A., Sweller, J., Kirschner, F., & Zambrano R., J. (2018). From cognitive load theory to collaborative cognitive load theory. *International Journal of Computer-Supported Collaborative Learning*, 13(2), 213–233.

Klepsch, M., & Seufert, T. (2020). Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load. *Instructional Science*, 48(1), 45–72.

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.

Lodge, J. M., Kennedy, G., Lockyer, L., Arguel, A., & Pachman, M. (2018). Understanding difficulties and resulting confusion in learning: An integrative review. *Frontiers in Education*, 3, 49.

Loibl, K., & Leuders, T. (2019). How to make failure productive: Fostering learning from errors through elaboration prompts. *Learning and Instruction*, 62, 1–10.

McDermott, K. B. (2020). Practicing retrieval facilitates learning. *Annual Review of Psychology*, 72, 609–633.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2), 81–97.

Nückles, M., Roelle, J., Glogger-Frey, I., Waldeyer, J., & Renkl, A. (2020). The self-regulation view in writing-to-learn: Using journal writing to optimize cognitive load in self-regulated learning. *Educational Psychology Review*, 32(4), 1089–1126.

Paas, F., & van Merriënboer, J. J. G. (2020). Principles of instructional design: Towards a cognitive load perspective on working memory use in education. *Applied Cognitive Psychology*, 34(4), 713–721.

Roediger, H. L., & Butler, A. C. (2010). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15(1), 20–27.

Sana, F., & Yan, V. X. (2022). Interleaving retrieval practice promotes science learning. *Psychological Science*, 33(5), 782–788.

Schneider, S., Beege, M., Nebel, S., Schnaubert, L., & Rey, G. D. (2021). The Cognitive-Affective-Social Theory of Learning in digital Environments (CASTLE). *Educational Psychology Review*, 33(1), 1–38.

Skulmowski, A., & Xu, K. M. (2021). Understanding cognitive load in digital and online learning: A new perspective on extraneous cognitive load. *Educational Psychology Review*, 33(4), 1473–1496.

Steenhof, N., Woods, N. N., Van Gerven, P. W. M., & Mylopoulos, M. (2019). Productive failure as an instructional approach to promote future learning. *Advances in Health Sciences Education*, 24(4), 739–749.

Sundararajan, N., & Adesope, O. O. (2020). Keep it coherent: A meta-analysis of the seductive details effect. *Educational Psychology Review*, 32(3), 707–734.

Sweller, J. (2023). The development of cognitive load theory: Replication crises and incorporation of other theories can lead to theory expansion. *Educational Psychology Review*, 35, 95.

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31(2), 261–292.

Tetzlaff, L., Schmiedek, F., & Brod, G. (2020). Developing personalized education: A dynamic framework. *Educational Psychology Review*, 32(3), 877–905.

Tetzlaff, L., Simonsmeier, B. A., Peters, T., & Brod, G. (2025). A cornerstone of adaptivity — A meta-analysis of the expertise reversal effect. *Learning and Instruction*, 97, 102142.

Weinstein, Y., Madan, C. R., & Sumeracki, M. A. (2018). Teaching the science of learning. *Cognitive Research: Principles and Implications*, 3, 2.