THE COGNITIVE FOUNDATIONS OF LEARNING

What We Know, What We Don't, and Where the Boundaries Are

Applied Pedagogy Research Lab

Guido Bartolucci, Principal Investigator

guido@appliedpedagogy.com

LAB.APPLIEDPEDAGOGY.COM

W2-001 · April 2026

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

CONTENTS

1

THE CLAIM AND ITS LIMITS

The cognitive science of learning has the strongest evidence base in all of education research. This claim — advanced by the Lo survey, confirmed by the L1 investigation, and unchallenged by any subsequent wave of this lab's work — is true. It is also, taken by itself, insufficient. The purpose of this review is to calibrate the claim: to establish how strong the evidence actually is, where it runs out, and what a curriculum designer can and cannot build on it.

The strength is real. Retrieval practice — testing yourself on material rather than rereading it — is supported by hundreds of studies across materials, ages, and settings, and has been rated "high utility" by the most comprehensive review of learning techniques in the field (Dunlosky, Rawson, Marsh, Nathan & Willingham, 2013)○. Spaced practice is "one of the most general and robust effects from across the entire history of experimental research on learning and memory" (Bjork & Bjork, 2011, p. 59)●. Cognitive load theory, the dominant framework for instructional design, rests on a cognitive architecture — limited working memory, expansive long-term memory, schema-based processing — that has been replicated across research centres for four decades (Sweller, van Merriënboer & Paas, 2019)●. The expertise reversal effect — the finding that instructional techniques helping novices can harm experts — has been confirmed across 60 studies with 5,924 participants (Tetzlaff, Simonsmeier, Peters & Brod, 2025)○. These are not fragile findings awaiting replication. They describe how human cognition processes information, and they hold because they reflect the architecture of the brain.

The limits are equally real. Nearly all of this evidence comes from well-structured domains — mathematics, physics, foreign- language vocabulary, medical diagnosis — where problems have clear correct answers and performance can be unambiguously measured. The extension to ill-structured domains — essay writing, historical reasoning, ethical analysis, design — is almost entirely untested. The cognitive science literature is overwhelmingly WEIRD: Western, Educated, Industrialized, Rich, and Democratic. Only 6% of classroom retrieval-practice experiments come from non-WEIRD countries (Murphy, Little & Bjork, 2023)●. The findings describe basic cognitive mechanisms that should in principle be universal, but the instructional prescriptions derived from them have been validated in a narrow slice of the world's educational contexts.

Most importantly, the cognitive science of learning treats cognition as if it operates in a motivational vacuum. Working memory limits, schema construction, retrieval strengthening — these mechanisms function the same way whether the learner is engaged or bored, safe or anxious, curious or compliant. But learners are not disembodied information processors. They are people with goals, emotions, social contexts, and histories. W2-008's curriculum review established that the capacity for warm, reciprocal relationship is the strongest predictor of adult flourishing, and that self-regulation develops primarily through warm, predictable environments rather than direct cognitive training (Watts, Duncan & Quan, 2018, cited in W2-008)○. The cognitive toolkit is necessary but not sufficient, and the affective-environmental envelope is prior. A student who has mastered retrieval practice but sits in a classroom where failure is punished will not use it. A spacing algorithm that optimizes retention intervals is useless for a learner who has stopped caring.

W2-009's competence review clarified what the cognitive toolkit enables: it builds Layers 1 and 2 of the competence stack — domain knowledge and procedural skill. These are the foundation, but they are not competence. The surgeon who knows anatomy and can execute procedures but

cannot judge when surgery is the wrong option is not competent. The transition from skill to judgment — Layer 3 — requires the qualitative reorganization of mental representations from surface features to deep structure (Chi, Feltovich & Glaser, 1981)•, a process the cognitive science literature describes but does not fully explain. What cognitive science offers is the raw material for competence; what it does not offer is a complete account of how that raw material becomes the capacity for judgment, self-monitoring, and epistemic honesty that defines genuine expertise.

This review proceeds through the major findings in order of evidential strength, engaging the primary texts that v1 cited from training knowledge and that v2 has now read directly. Every claim carries a provenance tag. Claims marked *Verified (direct)* rest on papers or books read in full during this investigation. Claims marked *Verified (via PI summary)* rest on detailed book summaries produced by the lab's reading-guide pipeline. Claims marked *Abstract-verified* rest on abstracts, metadata, or TLDRs. Claims marked *Training-derived* come from the agent's training knowledge and have not been verified against the source text. The provenance breakdown is reported in 'changelog.md'.

2 COGNITIVE LOAD THEORY

2.1 THE ARCHITECTURE

Cognitive load theory rests on a cognitive architecture that is by now uncontroversial in its broad outlines. Working memory is severely capacity-limited — approximately four chunks for novel information (Cowan, 2013)○ — and limited in duration. Long-term memory is vast, with no known limits on capacity or duration. Learning is the construction of schemas in long-term memory that can be retrieved and applied, chunking multiple elements into single units and thereby expanding the effective capacity of working memory. Expertise is not a matter of superior working memory but of superior long-term memory organization: experts possess large numbers of domain-specific schemas that allow them to process complex information as familiar patterns rather than as novel elements requiring individual attention.

Willingham (2021)◑ provides the clearest statement of the practical consequence: "Background knowledge allows chunking, which makes more room in working memory, which makes it easier to relate ideas, and therefore to comprehend" (§2.7.2). The baseball study (Recht & Leslie, 1988) demonstrated that students with baseball knowledge comprehended baseball-themed passages better than non-experts regardless of standardized reading level — knowledge status dominated reading skill as a predictor of comprehension. This is not a minor finding. It means that the capacity to think about something is not a general ability but a domain-specific consequence of what one already knows. "Memory is the residue of thought" — students remember whatever they actually think about during an activity, not what teachers intend them to learn (Willingham, 2021, §3.6.7)●.

2.2 THE THREE TYPES OF LOAD AND THE GERMANE REVISION

Cognitive load theory distinguishes three types of load on working memory during learning:

Intrinsic load arises from the inherent complexity of the material — specifically, from the number of elements that must be processed simultaneously in working memory (element interactivity). A simple vocabulary pair has low element interactivity; a multi-step physics problem where every variable depends on every other has high element interactivity. Intrinsic load cannot be reduced without changing what is being taught.

Extraneous load arises from poor instructional design — split attention between text and diagrams, redundant information, incoherent presentation. Extraneous load consumes working memory without contributing to schema construction. Reducing extraneous load is the primary design lever CLT offers.

Germane load was originally conceived as a third, independent source — the cognitive effort devoted to schema construction itself. Sweller has been candid about this concept's failure: "the idea of germane cognitive load was not generated from data. It was developed because it seemed to be a plausible and interesting idea. Nevertheless, over the years it became clear that there were many extraneous and several intrinsic cognitive load effects being generated, but there were no germane cognitive load effects being generated" (Sweller, 2023, p. 5)●. Germane load is now understood as redistributive — working memory resources shifted from extraneous activities to activities relevant to learning — rather than as a third independent source (Sweller, van Merriënboer & Paas, 2019)●.

The practical consequence is minimal — instructional design still aims to reduce extraneous load and maximize engagement with intrinsic complexity — but the theoretical architecture has changed. Trainers and textbooks that present a three-additive-component model are citing an outdated version of CLT.

2.3 PRIMARY AND SECONDARY KNOWLEDGE

The most important theoretical addition to CLT since its founding is the distinction between biologically primary and biologically secondary knowledge, drawn from Geary (2008). Primary knowledge — spoken language, face recognition, general problem-solving, basic social cognition — is "acquired easily, automatically, and unconsciously simply by immersion in a suitable environment. It normally does not need to be explicitly taught" (Sweller, 2023, p. 3)•. Secondary knowledge — reading, writing, mathematics, scientific reasoning — "is generally not acquired easily, automatically, or unconsciously but rather needs conscious effort. Educational institutions were developed to teach secondary knowledge" (Sweller, 2023, p. 3)•.

CLT applies exclusively to secondary knowledge. This resolves the discovery-learning debate that consumed instructional design for decades: discovery works fine for primary knowledge (that is how we evolved to acquire it) but is typically inferior to explicit instruction for secondary knowledge (Sweller, 2023)•. The distinction also constrains claims about generic cognitive-skill training. Self-regulation, for instance, is flagged by Sweller, van Merriënboer and Paas (2019)•as biologically primary: "Currently, there is virtually no evidence that self-regulation is teachable as a domain-general procedure that will improve performance on far transfer tasks." This aligns with the Watts et al. (2018) reframing that W2-008 documented — self-regulation develops primarily through environmental conditions, not through direct cognitive training.

2.4 ELEMENT INTERACTIVITY AS THE MASTER VARIABLE

Element interactivity is not merely one variable among many in CLT; it is the variable that determines whether any CLT effect will appear, reverse, or be absent. "Element interactivity cannot be determined just by reference to the structure of the task. The expertise of learners also determines element interactivity" (Sweller, 2023, p. 13)•. A physics problem that presents high element interactivity to a novice (who must hold all variables in working memory simultaneously) may present low element interactivity to an expert (whose schemas chunk the variables into familiar patterns). Every CLT effect — worked example, split attention, redundancy, modality, expertise reversal — depends on element interactivity, and every effect has boundary conditions set by it.

This has a practical consequence that CLT's popularizers sometimes obscure: you cannot apply CLT prescriptions without knowing the learner's current knowledge state. A worked example that reduces extraneous load for one student may impose it on another. An integrated diagram that eliminates split attention for a novice may create redundancy for an expert. CLT is not a recipe; it is a framework that requires continuous diagnosis.

2.5 THE CLASSICAL EFFECTS

The major CLT effects have replicated across research centres and continue to produce consistent results:

The **worked-example effect** — novices learn more from studying worked examples than from solving equivalent problems — remains the strongest and most practically useful finding. It has been

replicated across mathematics, physics, electronics, and programming (Sweller, van Merriënboer & Paas, 2019)●.

The **split-attention effect** — learning suffers when learners must mentally integrate information from physically separated sources — motivates the practical recommendation to place labels directly on diagrams rather than in separate legends.

The **redundancy effect** — presenting the same information in multiple forms (diagram plus text that restates the diagram) harms learning rather than helping it — is counterintuitive but well-replicated.

The **modality effect** — combining visual and auditory presentation outperforms visual-only for high-complexity material — has been confirmed by meta-analysis (Ginns, 2005, cited in Sweller, van Merriënboer & Paas, 2019). It is modified by the **transient information effect**: long spoken segments impose their own load because auditory information is transient and cannot be re-inspected, partially reversing the modality advantage (Sweller, 2023)●.

2.6 THE ILL-STRUCTURED DOMAIN GAP

The most significant limitation of CLT is also its most consequential: the entire empirical base is drawn from well-structured domains. Sweller's 2023 self-presentation of CLT does not engage ill-structured domains — humanities, design, ethics — at all (Sweller, 2023)●. This is not an oversight; it reflects a genuine uncertainty about whether the framework's core concepts apply.

The difficulty is conceptual, not merely empirical. Element interactivity presupposes that "elements" can be identified and counted. In mathematics, elements are variables, operations, and relationships with clear boundaries. In essay writing or historical reasoning, what counts as an "element" is itself a judgment call. Van Drie and van Boxtel (2007, Abstract- verified) developed a six-component framework for historical reasoning — sourcing, contextualization, argumentation, substantive concepts, meta-concepts — that does not map onto CLT categories. Webster (2008)○ documented that architectural design education, built on Schön's reflective-practitioner model, has developed in near-complete isolation from CLT, with no cross-citation in either direction.

This does not mean cognitive principles are irrelevant in ill-structured domains. Working memory is still limited. Prior knowledge still matters. Retrieval still strengthens memory. But the specific instructional prescriptions derived from CLT — use worked examples, reduce split attention, fade scaffolding as expertise develops — have not been validated outside STEM and procedural domains. De Jong (2009)○ identified this as one of CLT's central open problems: the framework's measurement tools, type distinctions, and effect predictions may not generalize.

The German *Bildungswissenschaft* tradition offers a partial perspective on this gap. Klieme et al. (2003, Abstract- verified) introduced the *Kompetenzmodell* framework for national education standards, distinguishing sharply between *Bildung* (formation of the whole person) and *Kompetenz* (measurable cognitive capability). Anglo-American CLT produces *Kompetenz* findings — measurable gains in specific cognitive tasks. But the curriculum mission, as W2-008 established, requires *Bildung* — the development of judgment, character, and practical wisdom that transcends any specific cognitive task. Whether CLT's framework can contribute to *Bildung* goals or is structurally limited to *Kompetenz* outcomes is a question neither tradition has addressed.

The French *didactique* tradition adds another dimension. Bara and Tricot (2017)○ synthesized embodied-cognition theory with CLT, examining how sensorimotor experience interacts with cognitive load during symbol acquisition. Standard CLT treats the cognitive system as disembodied; Bara and Tricot open a line it ignores: physical manipulation during learning can reduce intrinsic

load by offloading representational work onto sensorimotor systems. This is a genuine gap in the standard CLT framework that the French tradition has identified.

2.7 COUNTER-EVIDENCE

CLT's most important internal counter-evidence is the germane load failure, which Sweller himself documented (Sweller, 2023)●. Externally, the strongest challenge comes from Skulmowski and Xu (2021)○, who argued that interactive media, immersion, and disfluency can induce extraneous load while simultaneously promoting motivation and learning — challenging the standard "minimize extraneous load" prescription. Their proposal for constructive alignment of load design with learning outcomes suggests that the relationship between load and learning is more nuanced than the simple "reduce extraneous, manage intrinsic" prescription implies.

De Jong (2009)○ identified additional open problems: measurement of cognitive load relies heavily on self-report, the three-type distinction is difficult to validate empirically, and the framework's predictions become ambiguous when multiple effects interact. These are real limitations that an honest treatment of CLT must acknowledge.

3

RETRIEVAL PRACTICE AND THE TESTING EFFECT

3.1 THE CORE FINDING

Of all the findings in the cognitive science of learning, retrieval practice has the strongest and broadest evidence base. The core result is simple: retrieving information from memory — testing yourself — produces better long-term retention than restudying the same information for matched time. This is not a subtle effect. It is large, robust, and generalizable.

Karpicke and Roediger (2008)○ produced the canonical demonstration. Students learned foreign-language vocabulary through study-test cycles. After reaching criterion, items were assigned to four conditions in a 2×2 design crossing continued study (yes/no) with continued testing (yes/no). At one week, continued studying had essentially no effect on retention, while continued testing produced a large positive effect. The condition with repeated testing but no further study substantially outperformed the condition with repeated study but no further testing. Additional exposure confers negligible benefit; additional retrieval confers large benefit.

Dunlosky et al. (2013)○ evaluated ten learning techniques against a four-dimension generalizability framework (learning conditions, student characteristics, materials, criterion tasks) and rated practice testing "high utility" — the highest rating, shared only with distributed practice. The evidence generalizes across word pairs, prose, science, medical knowledge, maps, ages from elementary through older adults, and criterion tasks from recall through comprehension.

Murphy, Little and Bjork (2023)● — the Bjork lab's most recent synthesis — identified two routes by which testing enhances learning: an indirect route (tests reveal metacognitive gaps, making learners aware of what they do and do not know) and a direct route (retrieval itself strengthens memory traces). They documented specific design recommendations with practical implications: frequent low-stakes cumulative exams outperform high-stakes infrequent exams; competitive multiple-choice alternatives are required for the retrieval benefit (trivially easy foils do not force genuine retrieval); and — a finding that reframes the role of testing in instruction — pretesting enhances encoding even when initial performance is near zero. Testing is not just a tool for after instruction; it is a tool for before instruction as well.

3.2 BOUNDARY CONDITIONS

The testing effect is not without limits, and an honest treatment must specify them.

Feedback. Testing without corrective feedback can consolidate errors. The benefit depends on feedback, especially for complex material (Dunlosky et al., 2013, Abstract-verified; Murphy, Little & Bjork, 2023)●.

Material complexity. The testing effect is strongest for simple paired-associate material — low element interactivity in CLT terms. Van Gog and Sweller (2015)○ argued that the effect attenuates or reverses for high-interactivity material where worked examples are more effective. This is an important interaction between two otherwise independent literatures.

Individual differences. De Lima and Buratto (2024)○ investigated individual-difference moderators and found no consistent patterns — methodological heterogeneity prevents firm conclusions. The honest answer to "who benefits most from retrieval practice?" is: we do not know yet.

Test anxiety. Test-anxious students may show attenuated benefits from retrieval practice, though the evidence is unsettled. Murphy, Little and Bjork (2023)● noted that collaborative testing reduces anxiety while preserving the retrieval benefit — a practical workaround.

WEIRD limitation. Only 6% of classroom retrieval-practice experiments are from non-WEIRD countries (Murphy, Little & Bjork, 2023)●. The basic memory mechanism should be universal — it reflects how retrieval modifies memory traces — but the classroom-implementation findings may not generalize to educational systems with different norms, structures, and power dynamics.

Far transfer. The retrieval-practice advantage is largest for recall and near transfer. Evidence that retrieval practice promotes far transfer is substantially weaker. Sana and Yan (2022)○ found that interleaving retrieval practice (mixing question types during testing) promotes science learning more effectively than either technique alone (d = 0.35 over blocked quizzing), suggesting that the transfer limitation may be partially addressable through retrieval design rather than retrieval per se.

3.3 THE METACOGNITIVE DISCONNECT

Perhaps the most troubling finding in the retrieval-practice literature is not about what works but about what students actually do. Karpicke, Butler and Roediger (2009, Training- derived) surveyed 177 college students and found that the most commonly reported study strategy was rereading notes or textbooks. Few reported using self-testing. Students' study strategies were poorly aligned with the evidence on what works.

Worse, Karpicke and Roediger (2008)○ found that students' predictions of their own performance were uncorrelated with actual performance. Students could not accurately predict which study condition would produce better delayed recall. This prediction-performance disconnect is the empirical foundation for Kirk-Johnson, Galla and Fraundorf's (2019)○ misinterpreted-effort hypothesis, which is treated in Section VIII.

The practitioner community has recognized and attempted to address this disconnect. Brown, Roediger and McDaniel (2014)○ — a practitioner book written by the researchers themselves — translated the testing effect into accessible language, framing the gap between student behavior and evidence-based practice as "illusions of competence." Lemov (2010/2021)○ operationalized retrieval practice as classroom techniques — "Do Now," "Cold Call," "Exit Ticket" — that eliminate reliance on student self-regulation by building retrieval into the structure of the lesson rather than leaving it to individual study habits.

4

SPACING AND INTERLEAVING

4.1 SPACING: THE MOST REPLICATED FINDING

Distributed practice — spacing study sessions over time rather than massing them into single sessions — is one of the most replicated findings in the history of experimental psychology. "The benefits of spacing on long-term retention, called the spacing effect, have been demonstrated for all manner of materials and tasks, types of learners (human and animal), and time scales; it is one of the most general and robust effects from across the entire history of experimental research on learning and memory" (Bjork & Bjork, 2011, p. 59)●. Dunlosky et al. (2013)○ rated it "high utility" alongside practice testing — the only two techniques to earn this rating.

The practical recommendation is straightforward: distribute practice across days and weeks rather than massing it into single sessions. Willingham (2021)◑ summarized the evidence for a practitioner audience: "If you pack lots of studying into a short period, you'll do okay on an immediate test, but you will forget quickly. If, on the other hand, you study in several sessions with delays between them, you may not do quite as well on the immediate test but, unlike the crammer, you'll remember the material longer after the test" (§5.8.3)●.

The optimal spacing interval depends on the retention interval — how long the learner needs to retain the material. Cepeda et al. (2008)○ proposed a 10–20% heuristic: the optimal spacing gap is approximately 10–20% of the desired retention interval. For a final exam one month away, spacing study sessions three to six days apart is approximately optimal. This heuristic is useful but rough — it is based on limited studies and should not be treated as a precise prescription (Dunlosky et al., 2013)○.

Dehaene (2020)◑ provided the neural mechanism: during slow-wave sleep, hippocampal place cells replay recently encoded information at approximately twenty times speed, consolidating it into long-term cortical storage. "Every night, our brain consolidates what it has learned during the day … while we sleep, our brain remains active; it runs a specific algorithm that replays the important events it recorded during the previous day and gradually transfers them into a more efficient compartment of our memory" (§13.2.1)●. Spacing study across days exploits this consolidation mechanism — each sleep interval allows replay and integration that massed study forfeits.

4.2 THE SPACING MECHANISM DEBATE

Chen, Castro-Alonso, Paas and Sweller (2017)○ proposed a CLT-internal explanation: cognitive effort depletes working memory resources, and rest allows recovery. On this account, spacing works not because of forgetting-based retrieval benefits but because massed practice produces progressively depleted working memory, reducing learning efficiency. This depletion-recovery mechanism is compatible with the standard forgetting-based accounts but offers a different causal pathway — and one with practical implications for how breaks should be structured during intensive learning sessions.

The practitioner spaced-repetition community — particularly Piotr Wozniak (SuperMemo) and the developers of the FSRS algorithm — has generated insights the academic literature has not engaged. Wozniak's "two-component model" of memory (stability + retrievability) predicts

review timing better than single-parameter models, and FSRS outperforms SM-2 (Anki's default algorithm) on prediction accuracy across large user datasets (Wozniak, various; Ye, 2023)[○]. The two-component model is practitioner-originated and has no direct academic analogue, despite its predictive success in production systems processing millions of reviews. This is a genuine case where practitioner knowledge exceeds research coverage. No academic study has compared algorithmic scheduling against human-chosen spacing in a controlled design — the practitioner community treats algorithmic superiority as established, but the academic literature has not tested it.

4.3 INTERLEAVING: THE DISCRIMINATION EFFECT

Interleaving — mixing different types of practice problems within a study session rather than grouping them by type (blocking) — produces a moderate but reliable advantage for long-term retention and discrimination. Brunmair and Richter (2019)[○] conducted a multilevel meta-analysis of 59 studies with 238 effect sizes and found an overall interleaving advantage of Hedges' $g = 0.42$.

The mechanism is not spacing. Chen, Paas and Sweller (2021)[○] argued that spacing and interleaving are mechanistically distinct: spacing is a cognitive-load-recovery effect, while interleaving is a discriminative-contrast effect. They should not be conflated in curriculum design. Brunmair and Richter's (2019) meta-regression supports this distinction: interleaving effects are strongest when between-category similarity is high (the comparison forced by juxtaposition has more work to do), when within-category similarity is low (learners must extract deeper structural features), and when material is more complex. If interleaving worked purely through spacing, similarity structure would be irrelevant.

Birnbaum, Kornell, Bjork and Bjork (2012)[○] separated the discrimination and retrieval accounts experimentally and found evidence for both, but discrimination accounted for more of the effect. The sequential attention theory (Carvalho & Goldstone, endorsed by Brunmair & Richter) provides the explanation: interleaving shifts attention toward between-category differences; blocking shifts attention toward within-category similarities. Neither is universally superior — the better strategy depends on what the learner needs to notice.

4.4 WHEN INTERLEAVING HURTS

The most important boundary condition is the word-learning reversal: interleaving *hurts* for vocabulary learning ($g = -0.39$, Brunmair & Richter, 2019)[○]. Blocking is superior when the goal is building within-category coherence rather than between-category discrimination. This reversal is not a minor qualification — it means that interleaving is not a universal strategy but a discrimination strategy, useful when the learning goal involves telling similar things apart and counterproductive when it involves building up understanding of a single thing.

Effects for expository texts were non-significant (Brunmair & Richter, 2019)[○], suggesting that interleaving may not suit hierarchical, sequential content where understanding builds cumulatively.

The interleaving literature also reveals a tension with the practitioner language-learning community. Spaced-repetition users interleave vocabulary successfully across thousands of items; the meta-analysis shows blocking is superior for vocabulary. The discrepancy may reflect different definitions of "interleaving" — the practitioner community interleaves across a vast number of categories with high temporal spacing, while laboratory studies interleave a small number of categories within a single session. This deserves investigation.

4.5 NON-ENGLISH PERSPECTIVES

The Swedish variation theory tradition — developed by Marton and engaged extensively with Chinese mathematics education — offers a distinct perspective on what interleaving accomplishes. Watson and Mason (2006)[○] analyzed how exercise design structures sense-making: whereas the Anglo interleaving literature varies problem types randomly to force discrimination, variation theory varies features *systematically and purposefully* to highlight mathematical structure. Pang and Marton (2005)[○] demonstrated that teachers can use deliberate patterns of variation and invariance to make critical features discernible — a principled account of *what* should vary, not just *that* variation helps. This is a complement CLT does not offer: CLT asks "how much information?"; variation theory asks "what must vary so students see what matters?"

5

DESIRABLE DIFFICULTIES AND PRODUCTIVE FAILURE

5.1 TWO FRAMEWORKS, NOT ONE

Desirable difficulties and productive failure are frequently discussed as though they are the same thing. They are not. They share a surface similarity — both involve making learning harder in ways that improve long-term outcomes — but they rest on different mechanisms, target different learning outcomes, and have different boundary conditions. Treating them as identical obscures distinctions that matter for instructional design.

5.2 DESIRABLE DIFFICULTIES: THE BJORK FRAMEWORK

Robert and Elizabeth Bjork's concept of desirable difficulties rests on a central insight about the unreliability of performance as a guide to learning: "conditions that make performance improve rapidly often fail to support long-term retention and transfer, whereas conditions that create challenges and slow the rate of apparent learning often optimize long-term retention and transfer" (Bjork & Bjork, 2011, p. 57)•.

The theoretical backbone is the two-strength model of memory: "Storage strength reflects how entrenched or interassociated a memory representation is with related knowledge and skills, whereas retrieval strength reflects the current activation or accessibility of that representation and is heavily influenced by factors such as situational cues and recency of study or exposure" (Bjork & Bjork, 2011, p. 58)•. The key dynamic is that storage strength retards forgetting and enhances relearning, while retrieval strength merely reflects current accessibility. Learners misinterpret current retrieval strength (fluency) as storage strength (durable learning), leading them to prefer strategies that maximize fluency (massed practice, blocked study, rereading) over strategies that maximize durability (spacing, interleaving, testing).

The framework identifies four core desirable difficulties: varying conditions of practice (studying in different contexts improves retrieval), spacing study sessions (treated above), interleaving practice (treated above), and generation and testing. On the last, the Bjorks are direct: "Retrieval, in effect, is a powerful 'memory modifier'" (Bjork & Bjork, 2011, p. 62)•, and "any time that you, as a learner, look up an answer or have somebody tell or show you something that you could, drawing on current cues and your past knowledge, generate instead, you rob yourself of a powerful learning opportunity" (p. 62)•.

The "desirable" qualifier is critical: "Desirable difficulties, versus the array of undesirable difficulties, are desirable because they trigger encoding and retrieval processes that support learning, comprehension, and remembering. If, however, the learner does not have the background knowledge or skills to respond to them successfully, they become undesirable difficulties" (Bjork & Bjork, 2011, p. 58)•. This directly constrains the framework's scope: a difficulty is only desirable when the learner has enough prior knowledge to engage with it productively. For a complete novice, all difficulty may be undesirable.

But is "desirable difficulties" a theory or a label? The four phenomena it bundles — spacing, interleaving, testing, generation — may have different mechanisms. Spacing works through forgetting-based retrieval strengthening (the two-strength model). Interleaving works through dis-

criminative contrast (Brunmair & Richter, 2019)○. Testing works through retrieval-based memory modification. Generation works through deeper initial encoding. Whether these share a common mechanism is the unifying-mechanism question that Section IX defers to L2-F.

5.3 PRODUCTIVE FAILURE: KAPUR'S PROGRAM

Manu Kapur's productive failure program makes a different and more specific claim. Productive failure is not a general endorsement of difficulty during learning; it is a specific instructional design: problem-solving first, then explicit instruction that assembles the knowledge activated by the struggle.

The empirical record is strong. Kapur's meta-analysis of more than 50 studies reporting more than 160 experimental comparisons found that "the relative effect of learning from Productive Failure was up to three times (that's 300%) that of learning from a good teacher for one year … more than 80% of the studies in the above meta-analysis were independent replications" (Kapur, 2024, §7.12.9–§7.12.10)●. Both methods produced similar procedural knowledge, but productive failure produced significantly stronger conceptual understanding and transfer. The transfer effect was stronger than the conceptual effect — a finding with direct implications for curriculum designers concerned about knowledge application.

The mechanism operates through four processes — the 4A framework:

Activation. Attempting to solve a problem before instruction activates relevant prior knowledge — both correct and incorrect. "Attempting to answer questions about some topic before learning improves learning even if the initial answers were incorrect" (Kapur, 2024, §9.4.5)●. This activation is broader and deeper than what occurs during passive instruction.

Awareness. The experience of failure creates awareness of knowledge gaps. VanLehn et al. found that tutor explanations produced learning only when students were at an impasse — "when students were not at an impasse, the tutor's explanations did not result in learning" (Kapur, 2024, §10.1.7)●. Failure makes the gap visible; without the gap, instruction has nothing to fill.

Affect. The Zeigarnik effect — unfinished tasks are remembered better — creates a need for closure that productive failure harnesses. "The 'unfinished business' of failure propels the learner toward achieving closure from instruction that follows" (Kapur, 2024, §11.1.11)●. Productive failure students remained more curious even after receiving instruction (§11.9.10)●. Negative emotions — confusion, frustration — can be productive when experienced in a safe, supportive context. This connects directly to W2-008's finding that the affective-environmental envelope is prior: productive failure requires psychological safety.

Assembly. The instruction phase following problem-solving assembles the learner's partial solutions and prior knowledge into coherent understanding. "What we are not good at is analyzing students' incorrect answers to see if there are elements in those answers, bits and pieces and components, that could be used as building blocks for helping them learn the correct concept" (Kapur, 2024, §12.1.13)●. The teacher's role is not to present a pre-planned lecture but to build on what students actually produced during the struggle phase.

5.4 THE BASIC KNOWLEDGE FALLACY

Kapur's most important conceptual contribution is 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.

> "What my results show is how one learns the foundational knowledge influences how well they understand it and can transfer it. The learning path matters. Two different ways of learning the same basics can result in significantly different understandings and transfer. Learning is path dependent." (Kapur, 2024, §7.13.2)●

This is a direct challenge to the CLT prescription that novices should receive worked examples before practice. CLT optimizes for learning efficiency — minimizing the time to acquire correct procedures. Kapur argues that efficiency at acquisition can come at the cost of understanding and transfer. Two learners who have mastered the same procedure through different methods may have very different capacities for conceptual understanding, because one learned the procedure through a process that also built conceptual structure, and the other did not.

5.5 BOUNDARY CONDITIONS OF PRODUCTIVE FAILURE

Prior knowledge. He, Fiorella and Lemons (2025)○ produced a finding that complicates both CLT and PF predictions. In two experiments with biology students, the results were counter-intuitive:

- Experiment 1 (N = 367, low prior knowledge): problem-solving first significantly outperformed instruction first on near transfer. PF wins with novices. - Experiment 2 (N = 138, higher prior knowledge): instruction first significantly outperformed problem-solving first. DI wins with more knowledgeable learners.

Neither experiment showed significant effects on far transfer.

This creates an interpretive puzzle. CLT predicts novices need explicit instruction; He et al. found the opposite. PF predicts problem-solving first is generally superior; He et al. found it depends on prior knowledge — but in the opposite direction from what the CLT-PF interaction would predict. The resolution may involve the distinction between topic-specific prior knowledge and general domain experience: intro biology students have enough general science experience to generate meaningful attempts (making PF productive) while lacking the specific topic knowledge that would make instruction redundant (preventing the expertise reversal).

Domain specificity. PF has been studied primarily in mathematics and physics. Steenhof et al. (2019, Abstract- verified) extended it to health sciences, and DeCaro et al. (2023)○ replicated it in synchronous online physics courses. Baumgartner et al. (2025)○ demonstrated significant exam improvements (d = 0.28–0.59) in university linear algebra. But replication in genuinely ill-structured domains remains absent.

Consolidation dependence. Without competent teacher-led consolidation — instruction that explicitly connects student attempts to canonical solutions — PF produces unproductive failure (Kapur, 2016, Training-derived; Loibl & Leuders, 2019)○. Failure alone is insufficient; the error must be made visible and reflected upon. This makes PF dependent on teacher expertise in a way that retrieval practice and spacing are not.

Long-term durability. Kapur's meta-analysis covers immediate and short-term post-tests. Whether the PF advantage persists over weeks, months, or years is essentially untested.

Collaborative origins. The founding PF study (Kapur & Kinzer, 2008)○ used collaborative groups in a computer-supported learning environment. Groups generate more diverse solution representations than individuals, which may be part of the activation mechanism. Whether PF is equally effective for individual learners is less thoroughly tested.

5.6 THE SCHWARTZ AND BRANSFORD DISTINCTION

Schwartz and Bransford's (1998)◦ "preparation for future learning" tradition is closely related to but distinct from productive failure. Students who analyzed contrasting cases before instruction produced more accurate transfer one week later than those who read about contrasts, summarized text, or analyzed cases without instruction. The mechanism is perceptual differentiation — noticing distinguishing features creates readiness to absorb their significance during subsequent explanation.

The distinction matters: Schwartz and Bransford use carefully curated contrasting cases (perceptual learning); Kapur uses genuinely open problems where students fail (activation + awareness). These are different mechanisms producing different kinds of readiness. Both require subsequent instruction — the "telling" is essential — but the cognitive preparation they produce is different: perceptual differentiation in one case, prior-knowledge activation and gap awareness in the other.

5.7 THE JAPANESE PRECEDENT

Japanese mathematics education provides ecological validity for productive failure that purely experimental studies cannot. Stigler and Hiebert's (1999)◦ TIMSS video analysis showed that Japanese teachers present unsolvable problems, let students struggle, and then use student solutions — correct and incorrect — for class discussion leading to conceptual understanding. This is productive failure avant la lettre. Kapur's 4A framework is essentially a theoretical account of what Japanese lesson study classrooms had been doing for decades.

Fujii (2018)◦ clarified why it works at scale: lesson study and Teaching Through Problems are "structurally interdependent — two wheels of a cart." The institutional infrastructure of lesson study — collaborative teacher development through years of shared observation and refinement — is what makes PF sustainable. Teachers have collaboratively refined problem selection and consolidation, compensating for student novice status with teacher expertise. Without equivalent infrastructure, Western classroom replication is likely to underperform laboratory results.

6

TRANSFER: THE FIELD'S DEEPEST UNRESOLVED PROBLEM

6.1 WHY TRANSFER IS THE CENTRAL QUESTION

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

The v1 review identified transfer as the field's most significant gap. This investigation confirms that assessment and deepens it through direct engagement with the primary text.

6.2 THE BARNETT AND CECI TAXONOMY

Barnett and Ceci (2002)• argued that a century of transfer research had failed to resolve the question of whether far transfer occurs because researchers were comparing apples and oranges — studies that differed on multiple unstated dimensions while being discussed as if they tested the same phenomenon.

Their taxonomy crosses two factors. Factor A concerns *what is transferred* along three dimensions: the specificity of the learned skill (from specific procedure to general heuristic), the performance change measured (speed, accuracy, approach), and the memory demands (from executing with a hint to spontaneously recognizing, recalling, and executing without any prompt). Factor B concerns *where and when* transfer occurs along six context dimensions: knowledge domain, physical context, temporal context, functional context, social context, and modality.

The spontaneity dimension is especially important. Detterman (1993), quoted by Barnett and Ceci, stated it bluntly: "Telling subjects to use a principle is not transfer. It is following instructions" (p. 620)•. If a student applies a principle only when prompted to do so, the transfer is in the prompt, not in the learner. Gick and Holyoak's classic study — in which only 30% of participants transferred a military analogy to a medical problem even after being told the two were related (Willingham, 2021, §4.1.7– §4.1.8)• — illustrates the severity of the spontaneity problem.

6.3 THE EMPTY CELL

Barnett and Ceci's most important finding is an absence. When well-known transfer studies are plotted against the three most educationally relevant context dimensions (knowledge domain, physical context, temporal context), "one aspect of Figure 6 is worth particular note — the rightmost cell is empty. We simply did not find any well-controlled studies testing transfer to a far domain, in a far physical context, and in a far temporal context … that cell is of particular relevance to the educational applications of transfer research" (Barnett & Ceci, 2002, p. 627)•.

What general education claims to produce — transfer from school to workplace, years later, in a different physical and social context, using a different modality — "is, as yet, largely unknown" (p. 632)•. Most demonstrations of "far transfer" are actually near on most dimensions. Gick and

Holyoak's classic study was far only on knowledge domain; physical context, temporal context, functional context, social context, and modality were all near.

The paper identified only three isolated demonstrations of transfer that is far on two or more context dimensions: Bahrick and Hall (1991) on math retention 30+ years later; Chen and Klahr (1999) on control-of-variables transfer to a new domain at seven months; and Fong, Krantz and Nisbett (1986) on statistical reasoning to a far domain in a far physical context (Barnett & Ceci, 2002, pp. 629–630)●.

6.4 THE HONEST ASSESSMENT

Barnett and Ceci did not resolve the transfer debate. Their contribution is diagnostic: "after a century of intense research activity on the topic of transfer, scholars are perhaps in no greater agreement than they were at its inception" (p. 634)●. And: "Estimation of a single effect size for far transfer is misguided in view of this complexity" (p. 612)●.

The practical consequence is sobering. If transfer is the standard used to justify educational investment, "then it must apply well beyond the environment of training — that is, far transfer is required. Finding evidence of transfer from today's math class to tomorrow's math class is not sufficient" (Barnett & Ceci, 2002, p. 619)●. But far transfer as maximally defined — across domain, physical context, temporal context, functional context, social context, and modality — has never been demonstrated in a well-controlled study. No study has tested transfer that was far on even a majority of dimensions.

6.5 NEGATIVE TRANSFER

Transfer is not always positive. Barnett and Ceci documented negative transfer: Luchins (1942) found that participants transferred an elaborate water-jug problem-solving procedure to a simpler problem where a direct solution existed — applying the wrong schema to a problem that did not require it. "There are cases in which participants actually perform worse on the transfer task" (Barnett & Ceci, 2002, p. 617)●. Negative transfer is a reminder that transfer operates by pattern matching, and pattern matching can go wrong.

6.6 WHAT PROMOTES TRANSFER

The strongest evidence for instructional approaches that promote transfer comes from productive failure. Kapur's meta-analysis found that PF reliably outperforms direct instruction on transfer tasks, with effect sizes up to three times those of direct instruction (Kapur, 2024)●. The mechanism is encoding variability: when learners generate multiple solution representations during the problem-solving phase, they build richer, more flexible schemas that can be accessed from multiple retrieval routes. This flexibility is precisely what transfer requires.

Interleaving promotes transfer within the category- discrimination paradigm — learning to tell similar things apart (Brunmair & Richter, 2019)○. Retrieval practice promotes retention-based transfer — maintaining access to information over time. But neither addresses the full breadth of transfer that education claims to produce.

Schwartz and Bransford's (1998)○ "preparation for future learning" framework reframes the question: rather than measuring transfer as immediate application, measure it as readiness to learn from new instruction. This changes the measurement paradigm and produces more optimistic

results — but it also lowers the bar. Readiness to learn with help is not the same as spontaneous application without help, and it is the latter that education ultimately needs.

6.7 IMPLICATIONS FOR CURRICULUM DESIGN

The honest current answer to "can far transfer be reliably trained?" is: no, not by any known method, at any scale, with sufficient evidence to justify the claim. Near transfer can be promoted through varied practice, interleaving, and productive failure. Far transfer remains education's great unsolved problem.

This does not mean education is futile — near transfer is genuinely valuable, and the accumulation of domain-specific knowledge and skills is worthwhile in itself. But it means that curriculum designers should not promise far transfer from any instructional approach, because the evidence does not support such promises. The assumption that studying mathematics improves general reasoning, or that learning programming improves problem-solving, or that education in one domain makes learners broadly smarter, is not supported by the evidence. Transfer follows the contours of shared structure between tasks, and the amount of transfer decreases rapidly as structural similarity decreases. Willingham (2021)◑ states the implication directly: "The background knowledge that seems applicable almost always concerns the surface structure … That's why transfer is so poor" (§4.4.2)●.

7

EXPERTISE AND THE SKILL-TO-JUDGMENT TRANSITION

7.1 THE EXPERTISE REVERSAL EFFECT

The expertise reversal effect is not merely a boundary condition on the worked-example effect. It is a general principle that applies to all instructional techniques reviewed in this investigation: any technique that provides guidance the learner's existing schemas could have supplied becomes extraneous load for that learner. The effect was named and synthesized by Kalyuga, Ayres, Chandler and Sweller (2003)○, and it has since been confirmed as one of the most robust findings in instructional-design research.

Tetzlaff, Simonsmeier, Peters and Brod (2025, Abstract- verified) conducted the definitive meta-analysis: 176 effect sizes from 60 studies with 5,924 participants. The crossover pattern is crucially asymmetric:

- Low prior knowledge learners benefit from high-assistance instruction: d = 0.505 - High prior knowledge learners benefit from low-assistance instruction: d = −0.428

The asymmetry has direct design implications. The cost of over-assisting experts (d = −0.428) is real but smaller in absolute magnitude than the benefit of assisting novices (d = 0.505). When uncertain about a learner's level, erring toward more assistance is the less costly error — though not a costless one.

Three moderators reached significance. First, how prior knowledge is assessed matters — studies operationalizing expertise differently produce different ERE magnitudes. Second, the ERE is clearer in university students than in K–12 populations, where the evidence is "less clear." Third, and most importantly for this review: the ERE is robust in STEM and procedural domains but weaker in humanities and language learning. This domain moderator connects directly to the ill-structured-domain gap. If expertise in humanities operates differently from expertise in mathematics — as interpretive sophistication rather than as schema-based chunking — then the CLT mechanism that generates the expertise reversal effect may not apply in the same way.

A further qualification comes from Lachner, Russ, Hübner, Sibley and Scheiter (2025)●, who conducted an individual-participant-data meta-analysis of non-interactive teaching. Previously reported prior-knowledge aptitude- treatment interactions failed to replicate at large scale (N = 1,074 secondary physics students). The ERE may not generalize uniformly to all generative activities. Their finding that interest — not prior knowledge — moderated the retention effect connects to the cognitive-motivational integration treated in Section VIII.

7.2 SURFACE TO DEEP: HOW REPRESENTATIONS REORGANIZE

The cognitive mechanism underlying the expertise reversal effect — and underlying the transition from skill to judgment more broadly — is the reorganization of mental representations from surface-feature encoding to deep-structural encoding.

Chi, Feltovich and Glaser (1981)● produced the foundational demonstration. In their Study 1, eight experts (advanced physics Ph.D. students) and eight novices (students who had completed one semester of mechanics) sorted 24 textbook problems into categories. The results were qualitative, not merely quantitative:

Experts categorized by underlying physics principles — conservation of energy, Newton's second law, work-energy theorem. Novices categorized by surface features — inclined planes, pulleys, springs, blocks. Only five of twenty categories were shared between the two groups. This is not a difference in how much the groups knew; it is a difference in how their knowledge was organized. As Willingham (2021)◑ summarized: "It's not just that there is a lot of information in an expert's long-term memory; it's also that the information in that memory is organized differently … Experts don't think in terms of surface features, as novices do; they think in terms of functions, or deep structure" (§6.0.14)●.

Chi et al.'s (1981)● Study 2 revealed that expert categories contained procedural knowledge — not just what the problem type is but how to solve it and under what conditions the approach applies. Novice categories lacked this procedural component. Expert schemata encode solution methods with explicit conditions for applicability; novice schemata encode problem appearances without solution methods.

This reorganization is not accumulation — it is restructuring. The expert does not simply know more than the novice; the expert's knowledge is organized around different principles. This has a direct implication for instruction: teaching more facts does not automatically produce the reorganization that characterizes expertise. Something about the process of developing expertise causes the representational shift, and that process involves extensive engagement with varied, meaningful problems — not mere exposure to additional information.

Hmelo-Silver and Pfeffer (2004)○ extended this finding to complex biological systems: novices represented systems via visible components, while experts integrated structure, behavior, and function into coherent dynamic models. The surface-to-deep reorganization generalizes beyond physics, though it takes different forms in different domains.

7.3 THE TACIT DIMENSION

Polanyi's *The Tacit Dimension* (1966)◑ provides the philosophical foundation for why the skill-to-judgment transition cannot be fully captured by explicit instruction. "We can know more than we can tell" (§6.0.7)● is not merely a statement about the limits of verbal expression. Polanyi demonstrated that tacit knowledge operates through a proximal-distal structure: we attend *from* subsidiary particulars (which we cannot fully specify) *to* their focal meaning (which we can apprehend). The expert does not apply knowledge; they dwell in it and attend through it to the situation at hand.

The probe example illustrates the mechanism: "as we learn to use a probe … our awareness of its impact on our hand is transformed into a sense of its point touching the objects we are exploring" (§6.0.24)●. The tool becomes an extension of the knower. This is indwelling — and it maps directly onto the worked-example-to-fading transition: the learner must interiorize the example until it shifts from being an object of attention to a proximal term from which the learner attends to new problems.

Polanyi's warning about excessive formalization is relevant to CLT's limitations: "Scrutinize closely the particulars of a comprehensive entity and their meaning is effaced, our conception of the entity is destroyed" (§6.0.36)●. Over-scaffolding — providing too much explicit support for too long — may prevent the tacit integration that produces genuine understanding. This connects to the expertise reversal effect from a different theoretical tradition: what CLT explains as redundancy-induced extraneous load, Polanyi explains as the destruction of meaning through unbridled lucidity.

7.4 THE CONDITIONS FOR VALID EXPERTISE

Not all experience produces valid expertise. Kahneman (2011)[◑] documented the conditions — drawn from his adversarial collaboration with Gary Klein — under which expert intuition can be trusted:

> "If the environment is sufficiently regular and if the judge has had a chance to learn its regularities, the associative machinery will recognize situations and generate quick and accurate predictions and decisions. You can trust someone's intuitions if these conditions are met." (§29.0.32)[●]

Two conditions must hold: environmental regularity (the domain must contain stable, learnable patterns) and adequate feedback opportunity (the practitioner must receive clear, timely feedback on outcomes). Chess, firefighting, and many medical diagnoses meet these conditions. Stock markets, long-range political forecasting, and psychiatric prognosis do not.

This directly constrains the expertise reversal discussion. The ERE operates in domains where the Kahneman-Klein conditions are met — well-structured problems with clear feedback. In ill-structured domains where environmental regularities are weaker, expertise itself may be less reliable — which means the ERE may not apply in the same way. The domain moderator in Tetzlaff et al. (2025) — weaker ERE in humanities — is consistent with this: if expertise in humanities is less reliable to begin with, the reversal from "help novices" to "hinder experts" may be attenuated because the expert schemas are less robust.

The Kahneman-Klein conditions also constrain the deliberate-practice debate. Practice produces expertise most reliably in high-validity environments with good feedback. In low-validity environments — professions with delayed, noisy, or absent feedback — even thousands of hours of practice may not produce reliable expertise.

7.5 DELIBERATE PRACTICE: CONSTRAINING THE STRONG CLAIM

Macnamara, Hambrick and Oswald's (2014)[●] meta-analysis — 88 studies, 111 samples, 157 effect sizes, N = 11,135 — found that deliberate practice explains a variable and often modest fraction of performance variance:

Domain	Variance explained
Games	26%
Music	21%
Sports	18%
Education	4%
Professions	< 1%

The predictability of the domain environment moderates the relationship: high-predictability domains (running) show 24% variance explained; low-predictability domains (aviation emergency) show 4%. This aligns directly with the Kahneman-Klein conditions.

A methodological finding complicates the picture further. More valid measurement methods yield weaker practice- performance correlations. Retrospective interviews (asking people how much they practiced) yield 20% variance explained; log methods (contemporaneous records of actual practice) yield only 5% (Macnamara et al., 2014)[●]. The more valid the measurement, the weaker the relationship — a serious concern about recall bias inflating the practice-performance correlation.

The education finding — 4% variance explained, from 51 studies with 5,631 participants — is striking for curriculum designers. In educational contexts, factors other than accumulated practice time (prior knowledge, cognitive ability, motivation, instructional quality, self-regulation) dominate performance variance. Practice is "unquestionably important" (Macnamara et al., 2014)[●] but insufficient as a sole explanation.

Barth, Güllich, Macnamara and Hambrick (2022)● added a developmental finding: predictors of junior success are *opposite* predictors of senior world-class success. Early starts, specialized practice, and rapid initial progress predict junior performance but not adult expertise. Güllich, Barth, Macnamara and Hambrick (2023)● quantified the disparity: 89.2% of international-level juniors failed to reach international level as seniors; 82.0% of international seniors had not been international juniors. The two populations are 92.8% disparate. This is a direct challenge to the intuition that early performance predicts later excellence — and a curriculum-design warning against optimizing for short-term performance metrics.

7.6 LEARNING RATE UNIFORMITY

Koedinger, Carvalho, Liu and McLaughlin (2023)● reported a finding that reframes the individual-differences question. Across 1.3 million observations in 27 datasets (6,946 students; math, language, science), student learning rate is strikingly uniform: approximately 0.1 log odds per practice opportunity, with the difference between the 25th and 75th percentile learner amounting to roughly one extra practice opportunity needed. Variation in initial knowledge is an order of magnitude larger than variation in learning rate.

"Students do not show substantial differences in their rate of learning. These results suggest that educational achievement gaps come from differences in learning opportunities" (Koedinger et al., 2023)●. What looks like a "fast learner" is largely explained by higher initial knowledge, not higher learning rate. The main adaptive lever is quantity of practice opportunities, not fundamentally different instructional methods. Achievement gaps are reframed as opportunity gaps — an equity argument with direct practical implications that W2-008's normative framework would endorse.

The finding holds under favorable learning conditions (immediate feedback, tailored tasks, corrective instruction) but has not been tested under poor instructional conditions. Language learning showed higher learning-rate variation than math or science, possibly because vocabulary learning involves rote memorization of arbitrary mappings rather than systematically structured knowledge.

7.7 CONNECTION TO W2-009

W2-009's competence review documented the skill-to-judgment transition in detail — from Dreyfus's five-stage model through Klein's Recognition-Primed Decision making to the Kahneman-Klein conditions. This review does not duplicate that treatment. What the cognitive-foundations side adds is the mechanism: the surface-to-deep reorganization described by Chi et al. (1981) is the cognitive process that makes the Dreyfus transition possible. The worked-example-to-fading sequence is the instructional implementation of that transition. And the expertise reversal effect is the empirical evidence that the transition is real — that instruction must adapt because the learner's cognitive architecture changes as expertise develops.

The Russian activity-theory tradition offers a parallel account. Galperin's stepwise formation of mental actions (*planomernoe formirovanie*) — from material through verbal to internal stages — is structurally parallel to CLT's worked-example-to-fading transition, but was developed independently with no cross-citation (Rambusch, 2006; de Rezende & Valdes, 2006, Abstract- verified)○. Galperin adds a dimension CLT lacks: an account of *why* the stages must proceed in the order they do — from external-material through verbalized to internal — not just that worked examples should precede practice. This is the strongest candidate for a CLT–Russian activity theory bridge, and one that neither tradition has pursued.

8

COGNITIVE-MOTIVATIONAL INTEGRATION

8.1 THE GAP THAT MATTERS MOST

The v1 review identified the near-total absence of integration between cognitive science and motivational science as Gap 3 — the most practically consequential gap in the field. Cognitive science treats learning as computation: encoding, storage, retrieval, schema construction. Motivational science treats learning as goal-directed, emotionally laden activity: autonomy, competence, relatedness, interest, self-efficacy. Both perspectives have strong evidence. Neither alone accounts for how people actually learn.

This v2 review narrows the gap but does not close it. Three findings — Kirk-Johnson et al.'s (2019) misinterpreted-effort hypothesis, Fan et al.'s (2024) AI metacognitive laziness, and the Watts et al. (2018) reframing of self-regulation — together provide a sharper picture of how cognitive and motivational factors interact.

8.2 THE MISINTERPRETED-EFFORT HYPOTHESIS

Kirk-Johnson, Galla and Fraundorf (2019)○ documented a specific metacognitive error that explains why students avoid effective strategies: effort during learning is misinterpreted as a signal that the strategy is not working.

The causal chain: effort → perceived difficulty → perceived poor learning → strategy avoidance.

Four studies at the University of Pittsburgh confirmed the pattern. Students who used interleaved practice (Study 1) and retrieval practice (Studies 2–3) experienced more effort, interpreted that effort as evidence of poor learning, and preferred the less effortful alternative (blocked practice, rereading). Study 3 added a long-term retention test: students who chose retrieval practice performed better — directly contradicting their own judgment about what had worked.

This finding transforms the practical problem. The barrier to desirable difficulties is not student laziness or ignorance. It is a specific metacognitive inference error — a reasonable but incorrect interpretation of phenomenological evidence. Students are applying a heuristic (if it feels hard, it's not working) that is valid in most life contexts but systematically wrong for learning.

The implication is that informational interventions are insufficient. Telling students that retrieval practice works does not override the in-the-moment experience of effort-as- failure. What is needed is effort reattribution — explicit instruction that reframes the experience of difficulty as a signal of learning rather than a signal of failure.

Rea, Wang, Muenks and Yan (2022)○ provided partial counter-evidence that sharpens rather than undermines the finding: students *can* mostly recognize effective learning strategies when presented with descriptions. The gap is not knowledge but motivation — low self-efficacy, high perceived cost of effort, and entrenched habits. The problem is experiential, not declarative: students know what works when asked in the abstract, but the phenomenology of effort overrides that knowledge in practice.

Biwer, de Bruin and Persky (2022)● provided the strongest naturalistic evidence that strategy training can produce academic performance gains. Three 90-minute Study Smart sessions improved pharmacy students' metacognitive knowledge, decreased highlighting and rereading, and increased

interleaving and distributed practice. Bottom-quartile students showed the largest gains (from approximately 60% to 80% on exams), and in the trained cohort, rank differences substantially reduced — an equity finding. The limitation is serious (no random assignment, cohort comparison confounded by COVID-19), but the direction is clear: metacognitive interventions can shift student behavior from ineffective to effective strategies.

8.3 AI METACOGNITIVE LAZINESS

Fan et al. (2024)○ introduced a new threat to the cognitive-motivational integration: AI-assisted learners engage in fewer metacognitive processes. In a randomized lab experiment (N = 117 university students, academic writing task), the ChatGPT group showed significantly higher essay score improvement than other conditions but no significant knowledge gain or transfer advantage. AI inflated the product without improving the underlying knowledge structure.

The most troubling aspect is the mechanism: AI did not demotivate learners — motivation was equivalent across groups. The problem is process displacement, not motivational suppression. The learner feels fine, the product looks good, but the learning has not occurred. The specifically reduced metacognitive processes were evaluation, monitoring, and orientation (planning/goal-setting) — precisely the processes that desirable difficulties are supposed to engage.

Macnamara, Berber, Çavuşoğlu et al. (2024)● extended this analysis to expert performance. AI assistants that handle cognitively demanding subcomponents allow experts to continue performing at a high level while underlying cognitive skills go unpracticed. Because task-level performance remains high, the decay is invisible. "A surgeon using an AI assistant may believe their skills are still sharp because they have continued to perform operations at a high level … They may not consider how well they would be able to perform without the AI assistant" (Verified direct).

Three sub-mechanisms from Messeri and Crockett (2024, cited in Macnamara et al.) deepen the analysis: the illusion of explanatory depth (believing understanding is deeper than it is), the illusion of exploratory breadth (believing all possibilities have been considered), and the illusion of objectivity (failing to account for AI bias). These illusions mean the user *cannot detect* the cognitive offloading because AI-assisted performance is high and engagement feels complete.

Together, Fan et al. and Macnamara et al. form a convergent case: AI reduces cognitive engagement at acquisition (Fan), fails to maintain skill in experts (Macnamara), and the user is unlikely to notice (both). The design implication is that AI tools for learning should include *deliberate withdrawal* — intervals where AI withholds assistance to preserve the cognitive engagement that produces and maintains learning. The distinction that matters is between AI that *scaffolds* (temporary support enabling internalization, in Vygotsky's sense) and AI that *substitutes* (permanent external support replacing internal capacity).

The connection to cognitive load theory is precise: AI assistance may reduce extraneous load (good) while simultaneously eliminating the germane processing that builds schemas (bad). The practical question for AI-mediated learning tool design is not "how much can AI help?" but "which cognitive processes must the learner perform themselves for learning to occur?"

8.4 CURIOSITY AS METACOGNITIVE COMPUTATION

Dehaene (2020)◑ offered a framework for connecting cognition and motivation through curiosity: "Curiosity occurs whenever our brains detect a gap between what we already know and what we would like to know — a potential learning area. At any given moment, we choose, from the various actions that are accessible to us, those that are most likely to reduce this knowledge gap

… Curiosity would be the brain's governor, a regulator that seeks to maintain a certain learning pressure" (§11.5.2)●.

The Goldilocks zone — "between the boredom of the too simple and the repulsion of the too complex, our curiosity naturally directs us toward new and accessible fields" (§11.5.3)● — connects directly to the desirable-difficulties framework. If curiosity depends on accurate metacognitive monitoring of knowledge gaps, then metacognitive errors (confusing fluency with understanding, misinterpreting effort as failure) will misalign curiosity. Kirk-Johnson's finding is, on Dehaene's account, a disruption of the brain's curiosity governor — effort signals are misinterpreted, and the learner's curiosity is directed away from productive difficulty toward comfortable fluency.

8.5 THE ENVIRONMENTAL ENVELOPE

W2-008's curriculum review established a finding that reframes the cognitive-motivational integration: self-regulation develops primarily through warm, predictable environments rather than through direct cognitive training (Watts et al., 2018, cited in W2-008). The marshmallow test effect is approximately half the original size after controlling for family background. This is directly relevant to the v1 review's treatment of metacognitive and self-regulatory strategies: the ceiling on what direct cognitive training can achieve is probably lower than v1 implied.

Helm, Huber and Loisinger (2021)○ independently confirmed this from a different methodological tradition — large-scale survey of 255,955 students in Germany, Austria, and Switzerland during COVID-19 school closures. Self-regulatory success during closures was heavily mediated by family background and parental support, not by student- internal regulatory capacity.

The honest synthesis: the cognitive toolkit — retrieval practice, spacing, interleaving, productive failure — is necessary but can be undermined by motivational and environmental factors it does not control. A student who understands retrieval practice is effective but lacks the motivation to practice will not benefit. A student who is highly motivated but uses ineffective strategies will learn less than she should. And the intersection — how to get students to use effective strategies and to persist through the discomfort they produce — requires both cognitive and environmental interventions. Lemov's (2010/2021, Training- derived) observation that classroom management must precede instruction — that cognitive strategies cannot work in a chaotic environment — captures this: the environmental envelope is prior.

9 THE UNIFYING-MECHANISM QUESTION

9.1 WHAT L2-F MUST EXPLAIN

This review defers the investigation of predictive processing as a potential unifying framework to a dedicated L2-F agent. The purpose of this section is not to resolve the question but to sharpen the target: which specific findings would a unifying framework need to explain, and which findings resist unification?

The scorecard assembled during Session 2 identifies nine robust findings that any candidate unifying framework — predictive processing, schema theory, connectionist accounts, ACT-R, or any other — must accommodate:

1. **Testing effect.** Retrieval outperforms restudy. The unifier must explain why generation (with possible failure) outperforms exposure to the correct answer.

2. **Spacing effect.** Distributed practice outperforms massed practice. The unifier must explain why spacing helps even when the intermediate session is not a retrieval test — otherwise spacing is just a special case of the testing effect.

3. **Interleaving.** Interleaved practice outperforms blocked practice for discrimination. The unifier must explain the moderator: why does interleaving fail with too-dissimilar items or too-novice learners?

4. **Worked-example effect.** Passive observation of solutions produces learning for novices. A naïve prediction-error account would predict that worked examples should be inferior to attempted solving (because the latter generates prediction error), so the unifier must explain why passive processing can be effective.

5. **Expertise reversal effect.** What helps novices hurts experts. The unifier must specify what happens to its proposed learning signal as expertise grows.

6. **Productive failure.** Problem-solving before instruction outperforms instruction first for conceptual understanding. This is arguably the strongest case for a prediction-error unifier: the struggle phase generates prediction errors that make subsequent instruction more informative.

7. **Learning rate uniformity.** Under favorable conditions, learning rate is strikingly uniform across students (Koedinger et al., 2023). The unifier should predict this — a shared learning algorithm operating on different prior distributions would produce uniform rates with variable starting points.

8. **Forgetting curve.** Information decays approximately exponentially without retrieval. The unifier should make quantitative predictions about the decay curve, not just qualitative claims.

9. **AI metacognitive laziness.** AI that eliminates cognitive engagement eliminates learning (Fan et al., 2024; Macnamara et al., 2024). If the unifier's proposed learning signal is prediction error, then AI that eliminates prediction error should eliminate learning — which is what the evidence shows.

9.2 THE STRONGEST CASE FOR UNIFICATION

Three findings are most likely to be load-bearing for a predictive-processing unifier:

Productive failure maps naturally onto PP: the problem-solving phase generates high-entropy prediction errors that update the learner's generative model, making subsequent instruction (which supplies the correct model) more informative. Dehaene (2020)◑ makes the connection explicit: "The brain learns only if it perceives a gap between what it predicts and what it receives. No learning is possible without an error signal" (§12.1.2)●. This is the clearest bridge between cognitive science findings and predictive processing.

The expertise reversal effect maps onto amortized inference: the expert's model already predicts the solution path, so external scaffolding contributes zero new information and is processed as noise. The Tetzlaff et al. (2025) asymmetric-crossover finding quantifies what this would look like: a gradual reduction in prediction-error magnitude as schemas develop.

AI metacognitive laziness is the negative test: if prediction error drives learning, then AI that eliminates prediction error should eliminate learning. Fan et al. (2024) and Macnamara et al. (2024) confirm this prediction.

9.3 FINDINGS THAT RESIST UNIFICATION

Three findings resist easy incorporation:

Interleaving's mechanism — discrimination, not prediction error — does not map cleanly onto a prediction-error account. The similarity moderator (interleaving helps when categories are similar but discriminable) suggests the mechanism is attentional direction, not error correction.

The worked-example effect for novices — passive observation producing learning — requires a non-trivial PP claim: that the learner is internally simulating the solution as they read and generating prediction errors silently. This is plausible but not directly testable with current methods.

The spacing mechanism — if spacing works through WM-depletion-recovery (Chen et al., 2017) rather than through prediction error, then it is an architectural constraint, not a learning-signal phenomenon. A PP unifier would need to subsume the depletion account or treat spacing as operating through a different mechanism.

9.4 THE METHODOLOGICAL TEST

Hohwy and Seth (2020)○ provided the key methodological standard: the question is not "can PP explain X?" (it usually can, post hoc) but "does PP generate predictions about X that differ from those of schema theory, connectionist accounts, or generation-retrieval models?" If PP accommodates everything post hoc but predicts nothing differentially, it is a redescription, not an explanation.

The L2-F agent should look for differential predictions. The best candidates: (1) Does PP predict a specific shape for the spacing curve that differs from the Bjork two-strength prediction? (2) Does PP predict the exact crossover point in the expertise reversal effect from properties of the learner's generative model? (3) Does PP predict which specific metacognitive processes AI will displace (evaluation and monitoring, per Fan et al.) rather than making a generic "less prediction error = less learning" claim?

This review provides the findings; L2-F must provide the framework.

10 PRACTICAL IMPLICATIONS FOR CURRICULUM DESIGN

10.1 WHAT A CURRICULUM DESIGNER CAN CONFIDENTLY DO

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 (Karpicke et al., 2009)○. Build low-stakes quizzing, recall exercises, and retrieval opportunities into every unit. Include corrective feedback — testing without feedback can consolidate errors. Use pretesting before instruction as well as testing after it (Murphy, Little & Bjork, 2023)●. Vary the format of retrieval to match desired outcomes.

Space practice over time. Distribute practice across days and weeks. Build cumulative review into the schedule. When returning to previously covered material, use retrieval practice rather than re-exposition. The 10–20% heuristic provides rough guidance for spacing intervals (Dunlosky et al., 2013)○, but do not treat it as a precise prescription.

Use worked examples for novices, and fade them as expertise develops. For genuinely new material, provide fully worked examples that model the solution process. As learners develop competence, gradually remove steps (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. The asymmetry in the ERE (Tetzlaff et al., 2025)○ means that when uncertain about learner level, erring toward more support is the less costly error.

Interleave similar problem types during practice — but not all practice. When students are practicing skills that require discriminating between similar approaches, mix the problem types. This develops discriminative skill that blocked practice never exercises. But do not interleave vocabulary learning (where blocking is superior) or sequential concept building (where interleaving effects are non-significant) (Brunmair & Richter, 2019)○.

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 materials.

Explicitly teach students about effective learning strategies — including the phenomenology of effort. Students avoid desirable difficulties because effort is misinterpreted as poor learning (Kirk-Johnson et al., 2019, Abstract- verified). Informational interventions alone are insufficient; students need experiential evidence that effortful strategies produce better outcomes. Biwer et al.'s (2022)● Study Smart program provides a model.

10.2 WHAT A CURRICULUM DESIGNER SHOULD BE CAUTIOUS ABOUT

Productive failure requires design expertise and institutional support. PF is not "let students struggle and then teach normally." It requires carefully designed problems, competent teacher-led consolidation that builds on student solutions, and a classroom culture that supports risk-taking (Kapur, 2024)◑. Without these conditions, PF produces unproductive failure (Loibl & Leuders, 2019)○. The Japanese lesson study tradition (Stigler & Hiebert, 1999; Fujii, 2018, both Abstract-

verified) demonstrates that PF at scale requires institutional infrastructure for collaborative teacher development.

The prior-knowledge question in productive failure is unsettled. He, Fiorella and Lemons (2025)○ found that PF benefits depend on prior knowledge in complex and counter-intuitive ways. The simple recommendation "use PF for intermediate learners and DI for novices" may be wrong — but the right recommendation is not yet clear.

Extrapolation from STEM to other domains is underwarranted. Most evidence comes from mathematics, science, and technical domains. The expertise reversal effect is weaker in humanities and language learning (Tetzlaff et al., 2025)○. Historical reasoning has its own cognitive framework (van Drie & van Boxtel, 2007, Abstract- verified) that does not map onto CLT categories. The CLT and design-education literatures exist in near-complete parallel (Webster, 2008)○. Apply cognitive principles in ill-structured domains, but test and adapt rather than assume.

AI-mediated learning tools must preserve cognitive engagement. AI that performs cognitive work the learner should do undermines the learning it claims to support (Fan et al., 2024, Abstract-verified; Macnamara et al., 2024)●. Design AI tools that scaffold (temporary support for internalization) rather than substitute (permanent replacement of cognitive effort). Include deliberate withdrawal — intervals where AI withholds assistance. The Parasuraman and Manzey (2010) 70% accuracy threshold, cited by Macnamara et al., suggests that automation below 70% system accuracy produces worse outcomes than no automation at all.

10.3 WHAT THE EVIDENCE DOES NOT YET TELL US

Optimal spacing schedules for specific content. The 10–20% heuristic is rough. The practitioner community's algorithmic scheduling (FSRS, SuperMemo) has not been compared against human-chosen spacing in controlled academic studies. Curriculum designers should space practice but cannot specify optimal intervals with confidence.

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 no single correct answer exists. What constitutes a "worked example" in essay writing or historical analysis? What counts as "element interactivity" when elements are interpretive claims? These questions are unanswered.

How to promote far transfer reliably. Near transfer can be promoted through varied practice, interleaving, and productive failure. Far transfer — applying knowledge across genuinely different domains, contexts, and time scales — remains education's great unsolved problem (Barnett & Ceci, 2002)●.

How individual differences beyond prior knowledge interact with techniques. De Lima and Buratto (2024, Abstract- verified) found no consistent moderators. Working memory capacity, motivation, metacognitive skill, and personality have all been proposed as factors, but their relative importance and interactions are poorly characterized.

Long-term and cumulative effects. Most studies measure effects over days or weeks. Whether the benefits of retrieval practice, spacing, and productive failure compound over months and years is largely unknown. Curriculum designers need long-term data that the literature does not provide.

11

CLOSING ASSESSMENT

11.1 CONFIDENCE LEVELS

The following table summarizes the confidence level for each major finding reviewed, calibrated against the provenance standard:

| Finding | Confidence | Qualification | |———|———|———| | Retrieval practice enhances retention | High | Attenuates for high-complexity material | | Spacing enhances retention | High | Optimal schedules are rough estimates | | Interleaving enhances discrimination | Moderate–High | Reverses for vocabulary; non-significant for texts | | Worked examples benefit novices | High | Must be faded as expertise develops | | Expertise reversal effect | High | Weaker in humanities and K–12 | | Productive failure enhances conceptual understanding | Moderate–High | Domain-limited; consolidation-dependent | | Far transfer is reliably achievable | Low | Essentially untested at maximum definition | | CLT applies in ill-structured domains | Low | Virtually no evidence | | Cognitive-motivational integration | Emerging | Kirk-Johnson + Fan + Watts provide the pieces | | AI tools can undermine learning | Moderate | Theoretical + one strong empirical study | | Learning rate is uniform across students | Moderate | Under favorable conditions only | | Deliberate practice is the primary determinant of expertise | Low–Moderate | 4–26% variance; domain-dependent |

11.2 WHAT V2 RESOLVED

This v2 review advances beyond v1 in several specific ways:

Provenance. v1 cited Sweller, Bjork, Kapur, Chi, Willingham, Dehaene, Kahneman, and Polanyi primarily from training knowledge. v2 has engaged these sources at primary- text depth — either through full papers read via open access (Sweller 2019, Sweller 2023, Macnamara 2014, Chi et al. 1981, Barnett & Ceci 2002, Bjork & Bjork 2011) or through the lab's reading-guide pipeline (Willingham, Kapur, Dehaene, Kahneman, Polanyi). The provenance distribution is reported in 'changelog.md'.

The germane-load revision. v1 presented the three-type model as settled. v2 documents Sweller's own revision: germane load was "not generated from data" and is now redistributive, not additive (Sweller, 2023)●.

The primary/secondary knowledge distinction. v1 did not engage this. v2 documents it as CLT's most important theoretical addition since its founding, with direct implications for the discovery-learning debate and the limits of generic cognitive-skill training (Sweller, 2023; Sweller, van Merriënboer & Paas, 2019, both Verified direct).

The expertise reversal meta-analysis. v1 cited the ERE without quantitative precision. v2 anchors it in Tetzlaff et al.'s (2025) meta-analysis: d = 0.505 novice benefit, d = −0.428 expert cost, with domain and age moderators specified.

The He, Fiorella and Lemons (2025) finding. v1 flagged productive failure for genuine novices as Gap 7. v2 partially closes it: PF works *better* for low-prior-knowledge students, counter to CLT's novice prescription — but the far-transfer null limits the practical significance.

AI metacognitive laziness. This post-v1 finding (Fan et al., 2024; Macnamara et al., 2024) changes the design calculus for AI-mediated cognitive tools. v1 could not have engaged it.

Non-English traditions. v1 was entirely Anglophone. v2 engages German *Bildung/Kompetenz* (Klieme et al., 2003), French embodied-CLT (Bara & Tricot, 2017), Russian Galperin (Rambusch, 2006; de Rezende & Valdes, 2006), Japanese lesson study (Stigler & Hiebert, 1999; Fujii, 2018), and Swedish variation theory (Watson & Mason, 2006; Pang & Marton, 2005).

Practitioner gap detection. v2 includes systematic analysis of practitioner sources (spaced-repetition community, meta-learning authors, language-learning community, classroom practitioners) identifying six gaps where practitioner knowledge exceeds research coverage.

11.3 WHAT REMAINS GENUINELY UNKNOWN

Five questions remain open and are unlikely to be resolved by further literature review alone — they require new empirical work:

1. **CLT in ill-structured domains.** This was v1 Gap 1 and remains v2's most significant gap. The question of what constitutes element interactivity in interpretive domains is conceptual, not merely empirical.

2. **Far transfer mechanisms.** This was v1 Gap 2 and remains the field's deepest unresolved problem. Barnett and Ceci's (2002) taxonomy clarified the question; it did not answer it.

3. **Long-term cumulative effects.** v1 Gap 5 remains essentially untouched. The evidence base is dominated by studies measuring effects over days or weeks.

4. **Optimal spacing schedules.** v1 Gap 6 remains open. The practitioner community's algorithmic approaches (FSRS) have outpaced academic investigation, but no controlled comparison exists.

5. **Individual-difference moderators.** v1 Gap 4 has been narrowed only slightly. de Lima and Buratto (2024) confirmed the null rather than identifying moderators. Koedinger et al. (2023) reframed the question as one of opportunity rather than rate.

11.4 WHAT THE METHODOLOGY REVEALED

This domain is the hardest in the lab for the augmentation methodology to add value. The cognitive science evidence base is strong, well-replicated, and already well-reviewed. The easy wins available in other domains — finding stronger evidence, discovering replication failures, surfacing non-Western sources that overturn Western consensus — are less available here. Most major CLT, retrieval-practice, and spacing findings have replicated. The deliberate-practice contestation has stabilized. The transfer problem is genuinely unsolved, not merely under-investigated.

The value v2 adds over v1 is therefore different in character from the other W2 agents. It is not primarily about finding things v1 missed. It is about engaging primary texts at a depth that training knowledge cannot provide (Sweller's own revision of germane load, Kapur's basic-knowledge-fallacy argument, Barnett and Ceci's empty cell), about connecting the cognitive findings to the upper layers of the competence stack (W2-009's skill-to-judgment transition), about integrating the post-2022 AI evidence that changes the design calculus (Fan et al., Macnamara et al.), and about being honest about what the evidence does and does not establish.

The cognitive science of learning is genuinely the strongest part of the evidence base this lab investigates. But "strongest" is not "complete." The field's strength is in basic mechanisms — retrieval strengthens memory, working memory is limited, prior knowledge moderates instruction. These findings hold because they describe human cognitive architecture. The field's weakness is in prescription — in telling a curriculum designer exactly what to do for a specific learner learning

specific material in a specific context. The gap between mechanism and prescription is the gap that practitioners must bridge, and cognitive science provides the raw materials for that bridge without constructing it.

The most productive stance for a curriculum designer is not to treat these 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 and an affordance. Productive failure promotes transfer but 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. Cognitive science tells you the shape of the design space. It does not tell you the optimal point within it.

REFERENCES

Ahrens, S. (2017). *How to Take Smart Notes.*

Bara, F., & Tricot, A. (2017). [Embodied CLT for symbol acquisition.]

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.

Barth, M., et al. (2022). Predictors of junior versus senior elite performance are opposite. *Sports Medicine*, 52, 1399–1416.

Baumgartner, T., et al. (2025). PF design in university linear algebra.

Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2012). Why interleaving enhances inductive learning.

Biwer, F., de Bruin, A. B. H., & Persky, A. M. (2022). Study Smart. *Advances in Health Sciences Education*, 28, 147–167.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way. In *Psychology and the Real World*, pp. 56–64. Worth Publishers.

Brown, P. C., Roediger, H. L., & McDaniel, M. A. (2014). *Make It Stick.*

de Bruin, A. B. H., et al. (2020). Effort monitoring and regulation (EMR) framework.

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

Chen, O., Paas, F., & Sweller, J. (2021). Spacing and interleaving effects require distinct theoretical bases.

Chen, O., Castro-Alonso, J. C., Paas, F., & Sweller, J. (2017). Extending CLT to incorporate WM resource depletion.

Chen, O., Paas, F., & Sweller, J. (2023). Element interactivity as both definition and measure of task complexity.

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

Corcoran, A. W., Pezzulo, G., & Hohwy, J. (2020). From allostatic agents to counterfactual cognisers. *Biology & Philosophy*, 35.

Cowan, N. (2013). Working memory underpinnings of large-scale knowledge.

DeCaro, D. A., et al. (2023). Problem-solving before instruction in online physics.

Dehaene, S. (2020). *How We Learn: Why Brains Learn Better Than Any Machine … for Now.* Viking.

van Drie, J., & van Boxtel, C. (2007). Historical reasoning framework.

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

Edelsbrunner, P. A., et al. (2025). Beyond linear regression: Statistically modeling ATIs. *Learning and Individual Differences*, 102812.

Fan, Y., et al. (2024). Beware of metacognitive laziness. *British Journal of Educational Technology*.

Fiorella, L. (2023). Generative sense-making.

Forte, T. (2022). *Building a Second Brain.*

Fujii, T. (2018). Designing and adapting tasks in lesson planning. In *Mathematics Lesson Study Around the World* (ICME-13).

Ginns, P. (2005). Meta-analysis of the modality effect. *Learning and Instruction*, 15(4), 313–331.

Güllich, A., et al. (2023). Quantifying the extent to which successful juniors and seniors are disparate populations. *Sports Medicine*, 53, 1201–1217.

He, L., Fiorella, L., & Lemons, P. P. (2025). Does instruction-first or problem-solving-first depend on learners' prior knowledge? *Educational Psychology Review*.

Helm, C., Huber, S., & Loisinger, T. (2021). [COVID-19 teaching/learning in DACH countries.]

Hmelo-Silver, C. E., & Pfeffer, M. G. (2004). Comparing expert and novice understanding of complex systems.

Hohwy, J., & Seth, A. K. (2020). Predictive processing as a systematic basis for identifying the neural correlates of consciousness.

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

Kahneman, D. (2011). *Thinking, Fast and Slow.* Farrar, Straus and Giroux.

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

Kalyuga, S. (2025). Evolutionary perspective on human cognitive architecture. *Learning and Instruction*, 102300.

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

Kapur, M., & Kinzer, C. K. (2008). Productive failure in CSCL groups. *ijCSCL*, 3(3), 209–236.

Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success. *Educational Psychologist*, 51(2), 289–299.

Kapur, M., & Hattie, J. (2022). Fail, Flip, Fix, and Feed: A meta-analysis of flipped learning.

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

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning. *Memory*, 17(4), 471–479.

Kirk-Johnson, A., Galla, B. M., & Fraundorf, S. H. (2019). Perceiving effort as poor learning. *Cognitive Psychology*, 116, 101237.

Kirschner, P. A., Sweller, J., Kirschner, F., & Zambrano, J. (2018). From CLT to collaborative CLT.

Klieme, E., et al. (2003). [Competence model for German education standards.]

Koedinger, K. R., Carvalho, P. F., Liu, R., & McLaughlin, E. A. (2023). An astonishing regularity in student learning rate. *PNAS*, 120(13).

Koedinger, K. R., et al. (2012). Knowledge-Learning-Instruction (KLI) framework.

Lachner, A., et al. (2025). When does learning by non-interactive teaching work? *Educational Psychology Review*, 37, 88.

de Lima, R. H., & Buratto, L. G. (2024). Individual differences in retrieval practice.

Lodge, J. M., et al. (2018). Zones of optimal and sub-optimal confusion.

Loibl, K., & Leuders, T. (2019). How to make failure productive.

Macnamara, B. N., Hambrick, D. Z., & Oswald, F. L. (2014). Deliberate practice and performance in music, games, sports, education, and professions: A meta-analysis. *Psychological Science*, 25(8), 1608–1618.

Macnamara, B. N., et al. (2024). Does using AI assistance accelerate skill decay? *Cognitive Research: Principles and Implications*, 9, 46.

Martire, K. A., et al. (2025). Psychological insights for judging expertise. *Nature Reviews Psychology*, 4, 264–276.

Matuschak, A., & Nielsen, M. (2019). How can we develop transformative tools for thought?

Murphy, D. H., Little, J. L., & Bjork, E. L. (2023). The value of using tests in education. *Educational Psychology Review*, 35, 89.

Newport, C. (2016). *Deep Work.*

Oakley, B. (2014). *A Mind for Numbers.*

Pang, M. F., & Marton, F. (2005). Learning theory as teaching resource. *Instructional Science*, 33(2), 159–191.

De La Paz, S., et al. (2021). Strategy instruction in historical reasoning.

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

Rambusch, J. (2006). [Galperin's activity theory and situated learning.]

Rea, S. D., et al. (2022). Students can (mostly) recognize effective learning. *Journal of Intelligence*, 10(4), 127.

Recht, D. R., & Leslie, L. (1988). Effect of prior knowledge on good and poor readers' memory of text. *Journal of Educational Psychology*, 80(1), 16–20.

de Rezende, A., & Valdes, H. (2006). [Galperin's stage-formation theory.]

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning. *Psychological Science*, 17(3), 249–255.

Sana, F., & Yan, V. X. (2022). Interleaved retrieval practice promotes science learning.

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

Skulmowski, A., & Xu, K. M. (2021). Understanding cognitive load in digital and online learning.

Steenhof, N., et al. (2019). PF in health sciences education.

Stigler, J. W., & Hiebert, J. (1999). *The Teaching Gap.* Free Press.

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

Sweller, J. (2023). The development of cognitive load theory. *Educational Psychology Review*, 35, 95.

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*, 102142.

Tricot, A. (1998). [CLT review for French audience.]

Watson, A., & Mason, J. (2006). Variation and mathematical structure. *Mathematics Teaching*, 194, 3–5.

Webster, H. (2008). Architectural education after Schön.

Williams, A. M., & Hodges, N. J. (2023). Effective practice and instruction. *Journal of Sports Sciences*, 41(19), 1–18.

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

Wozniak, P. (various); Ye, J. (2023). FSRS algorithm.

Young, S. (2019). *Ultralearning.*